Deciphering Regulatory Networks in the Mouse Genome

Thesis

Deciphering regulatory networks in the mouse genome

Siddharth Sethi

Biocomputing
MRC Harwell Institute

Supervisor: Dr Ann-Marie Mallon
Dr Michelle Simon
Prof. Roger Cox

This dissertation is submitted for the degree of
Doctor of Philosophy

November 2019

Abstract

Regardless of all the major achievements in the field of genomics and in depth studies of the protein-coding genes, our knowledge about non-coding regions and their contribution in diseases remains incomplete. Large scale projects such as the ENCODE have produced a wealth of sequencing data which can be utilised to study epigenetic features associated with gene regulation. These studies have comprehensively identified regulatory elements such as enhancers in the human genome, but numerous questions still remain on their effect on gene function and disease causation.

The aim of this thesis is to identify enhancer regulatory networks in the mouse genome and investigate their effect on mouse models of human diseases. In order to study enhancer regulation, I have taken two approaches. First, I have produced a catalogue of well-defined multiple enhancer types in a diverse range of mouse tissues and cell-types. By systematically comparing different enhancer types, I found that super- and typical-enhancers have different effect on gene expression, but both are preferentially associated with relevant tissue-type phenotypes. Also genes associated with super- and typical-enhancers exhibit no difference in phenotype effect size or pleiotropy. Second, by utilising publicly available regulatory annotations, my enhancer catalogue and omics data, I have investigated regulatory mechanisms associated with metabolic and circadian mouse models. Here I identified novel regulatory networks or enhancers or transcription factor binding sites pertaining to the mutant mice.

In conclusion, my research has shown the usefulness of integrating enhancer annotations with an array of molecular data and has for the first time shown how different enhancer architectures influence gene function in the mouse genome. This study provides a valuable dataset to further characterise the mechanisms of gene regulation by enhancers in the mouse genome.

I would like to dedicate this thesis to my loving parents, my sisters, and my dear friend Paras Pathak who sadly is no longer with us.

Acknowledgements

This thesis has been a challenging journey and would not have been possible without the help and feedback of many people. First and foremost, I would like to thank my supervisors Dr Ann-Marie Mallon and Dr Michelle Simon for guiding throughout my DPhil. They supported me throughout this journey, helped me at every stage and challenged my limits to bring out the best in me. They always believed in my skills and provided me with great opportunities, for which I am truly grateful. Special thanks goes to Michelle; she was very patient while working with me and was always willing to help.

I would also like to to thank all members of the Mallon lab particularly Simon, John, Hugh, Luis and Henrik. A big thanks to Simon, who helped me with statistical problems, troubleshooting errors and writing more efficient algorithms. He was always there to install any computational packages I required on our server and modify the grid cluster according to my needs. John has been a great colleague and a friend. Thank you John for your help and support, and for teaching me the fundamentals of machine learning. Thanks to Hugh, Luis and Henrik who always provided me with critical feedback and were immensely helpful in my analysis of the IMPC data. I would also like to thank Ivan and Ilya from the Makeev lab, who guided me through the motif analysis of ChIP-seq data. Others I would like to thank include Prof Roger Cox and Dr Patrick Nolan for their fruitful collaborations and providing me the opportunity to be a part of their research.

I would like thank all my friends in Oxford and London, especially Jin, who was always there to help me with all my life problems. A big thanks to Miriam for being there during the last two years of my DPhil; you always brightened up my day after those long working hours, thank you. I would also like to thank Amy for proofreading my thesis; your corrections were really helpful.

Finally, I would like to thank my family for their unconditional love. Without their love and support, I would have not achieved anything in my life. A big thanks to my sisters for being the rock of my life.

Table of contents

List of figures

List of tables

Nomenclature

Acronyms / Abbreviations

3C Chromosome confirmation capture

4C Circular chromosome confirmation capture

5C Chromosome confirmation capture carbon copy

AD Alzheimer's disease

AML Acute myeloid leukaemia

ATAC-seq Assay for transposase accessible chromatin followed by sequencing

AUC Area under the curve

BAT Brown adipose tissue

BET Bromodomain and extra-terminal

Bmarrow Bone marrow

BmarrowDm Bone marrow derived marcophages

bp Base pairs

Brain_HM Brain hippocampus middle

Brain_ITL Brain inferior temporal lobe

Brain_MFL Brain mid frontal lobe

CAS9 CRISPR-associated protein 9

CDK Cyclin-dependent kinases

CH12 B-cell lymphoma cells

CHi-C Capture Hi-C

ChIP Chromatin immunoprecipitation

ChIP-seq Chromatin immunoprecipitation followed by sequencing

CKI Casein kinases

CRE *Cis*-regulatory element

CRISPR Clustered regularly interspaced short palindromic repeats

CRM *Cis*-regulatory module

DamID DNA adenine methyltransferase identification

DA-SNP Disease-associated SNP

DHS DNaseI hypersensitive site

DLBC Diffuse large B-cell lymphoma

DNase-seq DNaseI hypersensitive sites sequencing

ENU N-ethyl-N-nitrosourea

EPU Enhancer-promoter unit

eRNA Enhancer RNA

Esb4 Mouse embryonic stem cells

ESC Embryonic stem cell

Es-E14 Mouse embryonic stem cell line at day E14.5

ES Effect size

FAIRE-seq Formaldehyde-assisted identification of regulator elements followed by sequencing

GO Gene ontology

GWAS Genome-wide association study

H3K27ac Histone H3 lysine27 acetylation

H3K27me3 Histone H3 lysine27 trimethylation

H3K4me1 Histone H3 lysine4 monomethylation

H3K4me2 Histone H3 lysine4 dimethylation

H3K4me3 Histone H3 lysine4 trimethylation

H3K79me3 Histone H3 lysine79 trimethylation

HDL High-density lipoprotein

Heart_LV Heart left ventricle

Heart_RA Heart right atrium

Heart_RV Heart right ventricle

hESC Human embryonic stem cell

HSMM Human skeletal muscle myoblast

HUVEC Human umbilical vein endothelial cell

IGH immunoglobulin heavy chain

IMPC International mouse phenotyping consortium

kb Kilobase

LCR Locus control region

LD Linkage disequilibrium

LDL Low-density lipoprotein

Mb Megabase

MEF Mouse embryonic fibroblast

MEL Leukaemia cells, K562 analogue

MGD Mouse genome database

MNase-seq Micrococcal nuclease digestion followed by sequencing

MP Mammalian phenotype

MRCHI MRC Harwell Institute

mRNA Messenger RNA

NHEK Normal human epidermal keratinocyte

NHLF Normal human lung fibroblast

OCDEM Oxford Centre for Diabetes, Endocrinology and Metabolism

OR Odds ratio

PFM Positional frequency matrix

PIC Pre-initiation complex

PMCA Phylogenetic module complexity analysis

Pol II RNA polymerase II

PPI Protein-protein interaction

PR Precision-recall

PWM Positional weight matrix

ROC Receiver operating characteristic

RORE Retinoic acid-related orphan receptor response element

ROSE Rank ordering of super-enhancers

RPKM Reads per kilobase of transcript per million mapped reads

RRE *Rev-erbα*/ROR-binding element

Sci Short circuit

SCLC Small cell lung cancer

SCN Suprachiasmatic nucleus

SEC Super-enhancer class

SE Super-enhancer

SNP Single nucleotide polymorphism

SV40 Simian virus 40

T2D Type 2 diabetes

TAD Topologically associated domain

T-ALL T-cell acute lymphoblastic leukaemia

TEC Typical-enhancer class

TE Typical-enhancer

TFBS Transcription factor binding site

TF Transcription factor

tRNA Transfer RNA

TSRE Tissue-specific regulatory element

TSS Transcription start site

TTFL Transcriptional-translational feedback loop

Wbrain Whole brain

WEC Weak-enhancer class

ZT Zeitgeber time

Chapter 1

Background

Transcriptional regulation is a complex process which involves a complex network of transcription factors (TFs), co-factors and chromatin regulators binding to DNA regulatory elements like promoters and distal enhancers. Although precise control of gene expression is achieved with the help of multiple regulatory elements, enhancers play a central role in transcriptional regulation. Enhancers are capable of activating transcription of their target genes and driving cell-type specific gene expression which is important for the diversity in cell function. Genomic disruptions, either via genetic sequence variants or somatic mutations, within enhancer regions and the TFs associated with them, can cause mis-regulation of their target genes often leading to disease conditions. In this thesis, I will be studying gene regulation by enhancers to understand the disease conditions caused by their malfunction. In this chapter, I describe the concepts related to transcriptional regulation with a main focus on enhancers. I first provide a brief overview of how transcription is influenced by regulatory elements, followed by a comprehensive explanation of enhancers; their properties, interactions and chromatin regulators associated with their function. I then describe the current approaches in genomics to identify potential enhancers, followed by the identification of super-enhancers; dense clusters of active enhancers. Lastly, I explain how the malfunction of these regions can contribute to human diseases and traits.

1.1 Overview of mammalian transcriptional regulation

Cells are the basic structural and functional unit of living organisms. For the survival of an organism, the cells must respond to internal and external stimuli to perform the necessary functions. In a cell, thousands of genes are expressed which produce proteins and proteins control the cell function. Therefore, expression of genes in the correct amount is critical for optimum functioning of the cell and its related biological

processes such as cell differentiation and cell development. The synthesis of proteins from genes involves a series of tightly regulated steps, in which the genetic information required to synthesise the protein product flows from DNA to RNA (a process called transcription), followed by the conversion of RNA into a protein (a process called translation) (Crick, 1970; Crick et al., 1961). Transcription of a gene is controlled by regulatory elements such as its core promoter and distal enhancers. The core promoter of a gene is located just upstream of its transcription start site (TSS), while enhancers may be situated megabases (Mbs) away from the gene they regulate. Both core promoters and enhancers contain binding sites for TFs; proteins which recruit and interact with other co-factors, and collectively they activate or repress the rate of transcription process. The distal enhancers are brought into close proximity to the core promoter of a gene by the looping of the chromosome, which allows the enhancer-bound TFs to interact with the promoter-bound TFs and co-factors, and ultimately enhance the transcription process. The following section provides a brief overview of the concepts and processes associated with transcription, and how it is regulated by the regulatory elements such as promoters, enhancers and TFs.

1.1.1 Transcription

The first and the key regulatory step in gene expression is transcription, a process in which a segment of DNA is converted to its RNA copy by an enzyme called RNA polymerase II (pol II) (Fig. 1.1). Transcription initiates at a region immediate upstream of the TSS known as the core promoter. At the core promoter of a gene, a group of factors (together known as the pre-initiation complex (PIC)) required to initiate the transcription are assembled (Lenhard et al., 2012). The PIC comprises of pol II; general TFs; and co-factors such as the Mediator complex which are thought to bring TFs and co-activators close to the pol II. Once the PIC is assembled, pol II separates the two strands of DNA and uses one of the strands as a template to produce RNA. Pol II moves along the template DNA, reading one base at a time to produce a growing chain of complimentary RNA molecule in 5' to 3' direction (Kornberg, 2007). This process is known as elongation. Pol II generally transcribes 25-60 bases downstream before it pauses, a state which requires binding of additional factors to release pol II and continue its movement through the gene body. This pausing of pol II helps in the additional control of gene expression (recently reviewed in Mayer et al. (2017)). Once the complete gene is transcribed, the transcription is terminated by the signal from terminator sequences such as the chain of adenine nucleotides added at the 3' end of the RNA transcript by a process called polyadenylation (Connelly and Manley, 1988). After termination, the RNA transcript undergoes processing where the introns

are removed and exons are joined together (a process called RNA splicing) to form a single protein-coding sequence of mature messenger RNA (mRNA) (Wahl et al., 2009).

1.1.2 Translation

The mature mRNA produced from transcription is transported from the nucleus of the cell into its cytoplasm, where it comes in contact with the cell's protein production factory called the ribosome (Green and Noller, 1997) (Fig. 1.1). The ribosomes bind at the 5' end of the mRNA and move towards the 3' end. The process of translation initiates when the ribosomal units encounter a start codon (AUG) in the mRNA sequence. During this process, the sequence information in the mRNA is decoded into a chain of corresponding amino acids with the help of transfer RNAs (tRNAs) (Ramakrishnan, 2002). This chain of amino acids grows until the ribosomal complex encounters one of the stop codons (UAA, UAG, UGA) which marks the termination of translation and releases the complete amino acid chain. Overall, the amount of protein to be produced and its activity depends on the process of translation.

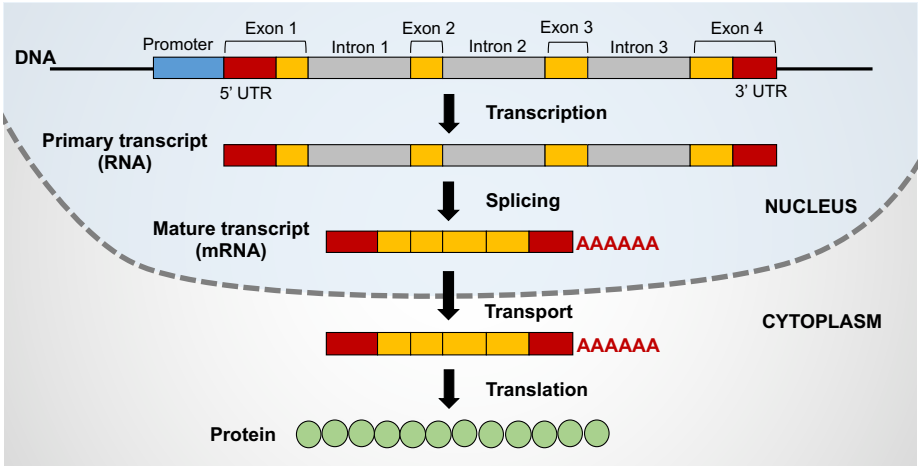


Fig. 1.1 Schematic of eukaryotic gene expression. The segment of DNA encoding for the gene comprises of exons and introns. This DNA segment is transcribed into a RNA molecule inside the nucleus, which undergoes processing such as splicing of introns and polyadenylation to produce the mature mRNA. The mRNA is then transported into the cytoplasm where its sequence is translated into a corresponding chain of amino acids and the protein product is released. Figure adapted from https://www.ncbi.nlm.nih.gov/probe/docs/applexpression.

In an organism, genes are tightly regulated i.e. they are expressed only when they are required to carry out a function in the body. In eukaryotes, transcription occurs within the nucleus while translation occurs in the cytoplasm, which provides a cell with an opportunity to regulate gene expression at every phase. For instance, a cell may control its protein production by (1) regulating gene transcription; (2) regulating splicing of RNA transcript; (3) regulating mRNA transport and localisation into cytoplasm; (4)

regulating which mRNAs undergo translation; (5) regulating the rate of translation; and (6) post-translational control (Darnell, 1982). However, for the majority of the cells, the most effective method to control the production of genes is at the transcriptional level as it decides whether a gene should be transcribed or not to produce the mRNA (Guenther et al., 2007a). Such control over transcription is mainly achieved by non-coding regulatory DNA sequences known as *cis*-regulatory elements (CREs), and their associated proteins such as TFs. The following sections describe the non-coding regulatory elements and TFs in detail.

1.1.3 Non-coding DNA

The non-coding DNA is the portion of the genome which does not contribute to a gene, and hence is not translated into a protein. These regions, previously alluded to as 'junk DNA', are scattered throughout the genome. They make up $\sim$90% or more of the genome in higher organisms. For instance, only $\sim$2% of the human genome is estimated to be protein-coding (approximately 20,000 - 25,000 genes) and the rest $\sim$98% of the genome is considered to be non-coding. Some non-coding regions can get transcribed, but do not undergo the process of translation and therefore, no functional protein is made. For this reason, the non-coding regions were initially considered to have no functional role in an organism. The term 'junk DNA' was first used by Susumu Ohno in 1972 to represent pseudogenes (defective DNA segments related to known genes, but do not code for proteins) (Ohno, 1972), but soon the term used was to represent all non-coding DNA sequences such as *cis*- and *trans*-regulatory elements, introns, non-coding functional RNA, repeat sequences, telomeres and transposons (Comings, 1972). After studying DNA segments across multiple species, Ohno hypothesised that the non-coding regions may have been functional in the past, but they have lost their usefulness through our evolution into more complex higher organisms (Ohno, 1972). This formed the notion that non-coding DNA is not under the effect of selective pressure, and genetic variations occurring in these regions would not disrupt any biological function.

Since non-coding DNA was considered to be functionally inactive, the majority of the biological research was conducted on the protein-coding part of the DNA. But as the human genome became more accessible, some researches discarded the notion that all non-coding DNA is junk (Kuska, 1998). With the completion of human genome in 2003, and the availability of next generation sequencing technologies, the ENCODE project (Birney et al., 2007) was the first systematic study to investigate the non-coding portion of the human genome using various state of the art sequencing techniques. The ENCODE pilot project inspected 1% of the human genome (or 30 million bases collectively), and discovered that some regions within the non-coding DNA sequence

are functional and play a critical role in controlling gene expression (Birney et al., 2007). The next phase of the ENCODE project inspected the complete human genome, which identified that up to 18% of the human DNA sequence is involved in controlling the 2% protein-coding portion of the human genome (ENCODE Project Consortium, 2012). Furthermore, they reported that approximately 80% of the human genome could be associated with some sort of biochemical function (ENCODE Project Consortium, 2012), hence scrapping the belief that the majority of the DNA is junk.

1.1.4 *Cis*-regulatory elements

Cis-regulatory elements (CREs) are one of the most important type of non-coding DNA sequences which facilitate binding sites for TFs and/or other regulatory molecules required to appropriately regulate the gene expression (Wittkopp and Kalay, 2011). The regulation of gene expression by CREs can involve either activating, repressing or sustaining expression levels. Gene expression is regulated by an interplay between five main types of CREs, categorised based on their function (Fig. 1.2):

1. **Promoters**: This is the stretch of DNA sequence $\leq$1 kilobase (kb) upstream of the TSS where the transcription initiates. The process of transcription activation requires a set of TFs to bind the promoter region in a particular order. Once this is achieved, pol II binds to the promoter and starts transcribing the gene downstream (Butler and Kadonaga, 2002). Most genes have a single promoter element close to the TSS, but some genes can contain multiple promoters to initiate transcription under certain circumstances.

2. **Enhancers**: These elements are segments of DNA, usually a few hundred base pairs (bp) long, which harbour binding sites for TFs that increase the basal level of transcription achieved by promoters. The TFs that bind to enhancers are known as transcription activators. Enhancers can be present upstream, downstream, within the introns or even several Mbs away from the gene they regulate (known as the target gene) (Kleinjan and Lettice, 2008). Their function is also independent of their orientation with respect to the target gene. A gene can be regulated by multiple enhancers, and likewise, one enhancer can regulate more than one gene.

3. **Silencers**: These elements harbour binding sites for TFs that inhibit the transcription of genes. The TFs that bind to silencers are known as repressors. The repressor proteins can either function independently or together with other co-repressors (Sertil et al., 2003). Similar to enhancers, silencer function is independent of its orientation and position with respect to its target gene (Ogbourne and Antalis, 1998). A classical example of a silencer element is the AT-rich *OCT1*

binding domain found in the promoter (∼140 bp upstream of TSS) of thyroid stimulating hormone beta gene (*TSHB*), which has been shown to repress the expression of *TSHB* (Kim et al., 1996).

4. **Insulators**: These elements protect genes from being inappropriately regulated by nearby regulatory elements. Insulators when located between an enhancer and a promoter, works as an interaction blocker between them, hence, preventing the enhancer to activate the transcription of its nearby gene (Kellum and Schedl, 1992; West et al., 2002). Insulators can also act as barriers to prevent irrelevant gene silencing by condensed chromatin in the surrounding regions (Sun and Elgin, 1999).

5. **Locus control regions (LCRs)**: A LCR is composed of multiple CREs which together influence the expression of a group of genes (Li et al., 2002). The elements within a LCR (for instance, enhancer and silencer elements) cooperatively exert a strong enhancer activity and drive the expression of their associated genes in a cell-type specific manner. An example of such a region is the well characterised LCR associated with the β-globin locus (Grosveld et al., 1987), located approximately 25 kb upstream of the β-globin genes. This LCR is essential for the expression of β-globin genes in erythroid cells.

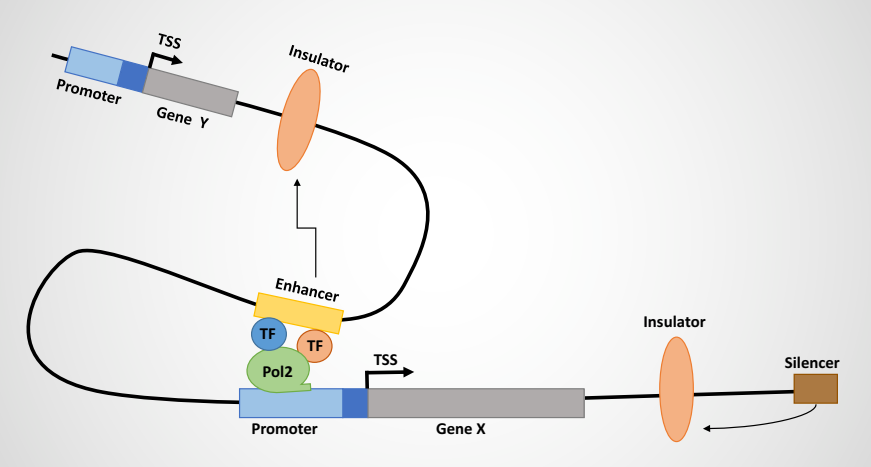


Fig. 1.2 Various types of *cis*-regulatory elements. A schematic showing the various types of *cis*-regulatory elements in the genome. Promoters are located just upstream of TSSs, and are composed of a core promoter (dark blue) and a proximal promoter region (light blue). Promoters assemble the basal transcriptional machinery to initiate transcription, while enhancers interact with promoters to increase the level of transcription. Silencers exhibit the opposite effect to enhancers by repressing the level of transcription. Insulators act as barriers to prevent the influence of enhancers and silencers on neighbouring genes. Figure adapted from Luizon and Ahituv (2015).

1.1.5 Transcription factors

The transcription of genes by pol II is facilitated by a family of regulatory proteins called TFs, and other factors capable of controlling the chromatin structure. TFs are the largest family of proteins in humans, forming approximately 9% of the genes (Babu et al., 2004). Together, these proteins help assemble the basal transcriptional machinery at the core promoter, hence playing a key role in transcription initiation. Apart from promoters, TFs also bind at enhancer regions where they have a central role in enhancer activation. TFs can be categorised into two types: general and sequence specific. General TFs mostly interact with pol II to initiate transcription, while the latter bind to specific targets to drive specific patterns of gene expression (Kadonaga, 2004). TFs perform their function by binding to the DNA either directly, or indirectly via another protein. The TFs capable of direct binding, however, can only bind at regions with specific sequence patterns, hence are referred to as sequence specific TFs. This sequence specificity of TFs is believed to be a result of chemical bonds and Van der Waals interactions between DNA binding domains of TFs and nucleotides in the DNA (Luscombe et al., 2001). Moreover, TFs usually bind in clusters within the enhancer regions in specific combinatorial binding patterns (Yan et al., 2013). The TFs involved in cooperative binding often have protein-protein interactions amongst themselves. Once the necessary TFs bind to the DNA, they can either activate or repress the expression of the associated gene. In eukaryotic organisms, the transcriptional TF network is complex as most of the genes require multiple TFs for their regulation, and a single TF typically can regulate many genes.

Transcription factor binding

DNA sequences functioning as enhancers contain short sequences (usually 6-12 bp) to which sequence-specific TFs bind. These short DNA sequences are known as motifs. At some positions in the motif sequence, the TFs are often able to bind to more than one nucleotide, which are referred to as ‘degenerate’ positions. Detecting motifs in the genome can help to identify the location of the binding sites of their associated TFs, and can also identify other co-factors binding near them, which could be involved in the same functional pathway. However, identification of genomic sites where TFs bind is a difficult problem and an ongoing challenge in the field of genomics. Two approaches have been commonly used to model the sequence binding specificity of a TF.

The first approach is the consensus sequence model, in which the sequence variability in the binding sites are summarised into a consensus sequence (Fig. 1.3A-B). The degenerate positions in the consensus sequence are represented by distinct symbols.

However, the consensus sequence fails to measure the variability in the nucleotides at each position.

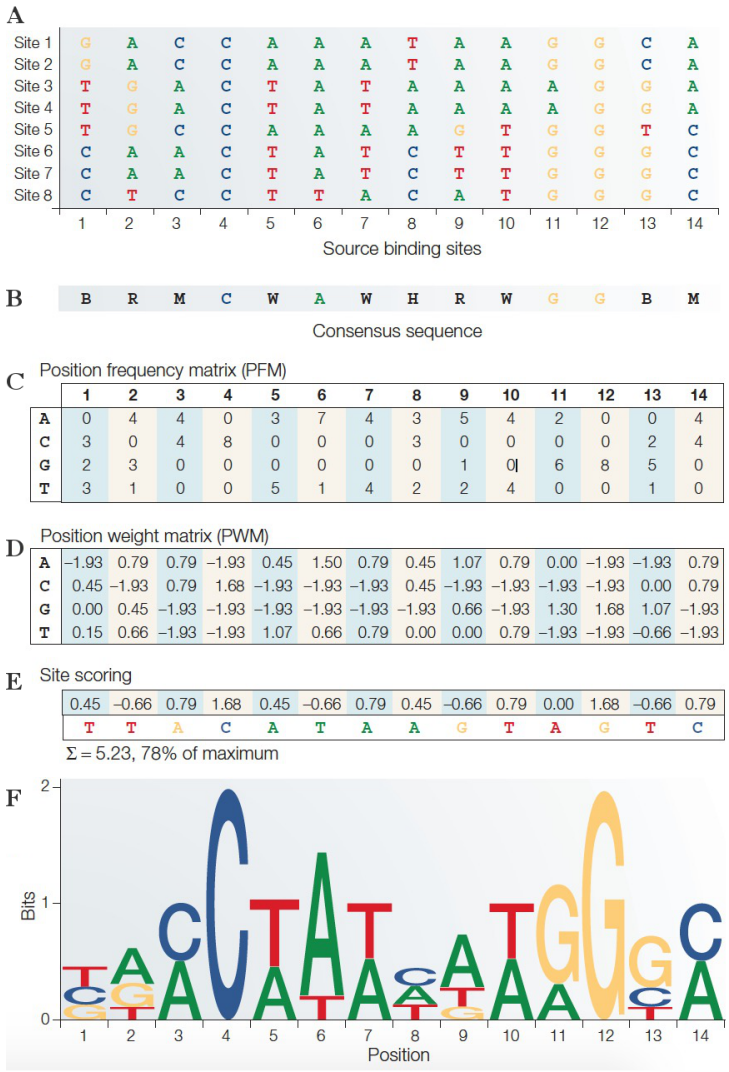


Fig. 1.3 Computational modelling of TF binding specificity. For a set of DNA sequences bound by a TF (A), the consensus sequence (B) summarises the variability in binding sites by representing patterns of nucleotide occurrences with distinct symbols. PFM (C) and PWM (D) on the other hand, more accurately quantifies the nucleotide variability at each position of the binding site, which can be used to calculate a score (E) representing the binding energy of the TF. The PWM of a TF can be visualised in a sequence logo (F) for easy interpretation. Figure taken from Wasserman and Sandelin (2004).

Second and a better approach is the position weight matrix (PWM) model (Stormo, 2000), which more accurately quantifies the nucleotide variability at each position. In this method, the number of observed nucleotides at each position of the binding site are first counted to construct a position frequency matrix (PFM) (Fig. 1.3C), and then normalised, converted to probabilities compared to background distribution and transformed to log-scale (Fig. 1.3D). Using this model, a score for a DNA sequence is calculated by summing the values in the PWM for each nucleotide (Fig. 1.3E), which

corresponds to the binding energy of a TF (Stormo, 2000). Finally, for easy and fast interpretation, the PWMs are visualised as sequence logos where the height of each nucleotide is scaled according to its observed frequency (Fig. 1.3F). To date, PWM is the preferred method to display the DNA binding preference of a TF.

1.1.6 Chromatin - a gatekeeper of regulatory regions

DNA is normally wrapped around histones which results in a compact configuration of dense nucleosomes inside the nucleus. However, researchers in the 1970s found that genomic locations associated with active genes are less compact compared to non-active genes (Axel et al., 1973; Weintraub and Groudine, 1976). More recent studies have identified TFs to bind DNA regions which are depleted of nucleosomes. For example, *in vivo* binding sites of TFs have been observed to strongly correlate with nucleosome free regions in *D.melanogaster* (Li et al., 2011) and mammalian genomes (John et al., 2011). This identified a critical role of chromatin as an accessibility barrier in TF binding and enhancer activity. It is now known that the dense nucleosome structure of the DNA does not allow TFs to bind, hence, this state is referred to as 'closed' or 'inaccessible' chromatin. On the other hand, when the region is nucleosome free, it is 'open' and accessible for TFs and other co-regulators to bind.

The conversion of a 'closed' chromatin state into 'open' chromatin, and vice-versa, is mostly controlled by specific regulatory proteins known as pioneer factors (Fig. 1.4). These proteins have the ability to disrupt the chromatin structure, which allows them to bind to the DNA even when it is inaccessible. By doing so, they incorporate chromatin-remodelling complexes which reposition the nucleosomes in this region and make it accessible to other TFs, thereby allowing the enhancer to assemble the required TF complex for its activation. This phenomenon was first discovered in yeast (Almer et al., 1986), where the activation of the *PHO5* gene was accompanied by the removal of nucleosomes at an upstream regulatory sequence. Soon after, the TF *HNF3* was discovered in the mouse to be essential for the activation of *Alb1* enhancer in liver (Liu et al., 1991). An example of a well studied pioneer factor is *PU.1* (also known as *SPI1*), which is essential for the generation of macrophages and B-cells (Scott et al., 1994a). To date, many pioneer TFs have been discovered such as *FOXA1* (Cirillo et al., 2002), *MYOD1* (Serna et al., 2005) and *PAX5* (McManus et al., 2011), which play an essential role before the regulatory networks initiate. Depending on the TF function and requirement, a pioneer factor may bind an enhancer for a short span until the enhancer becomes active (Hoogenkamp et al., 2009; Liber et al., 2010), or in some scenarios, may remain enhancer-bound and drive cell-type specific lineages (Mercer et al., 2011).

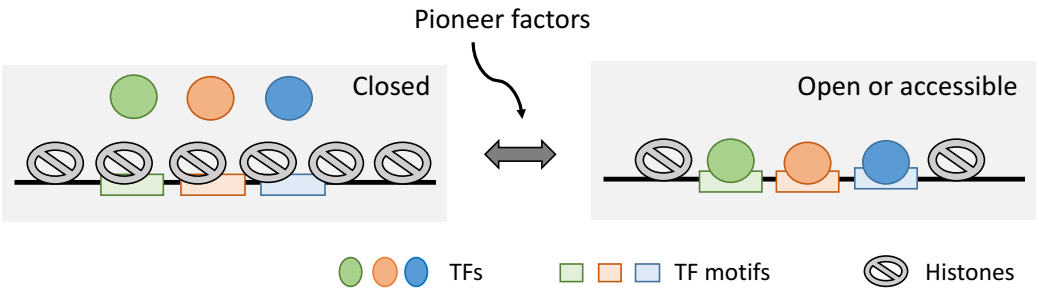


Fig. 1.4 Chromatin accessibility and pioneer factors. Chromatin acts as an accessibility barrier between TFs and enhancer regions. DNA regions with a high density of nucleosomes can restrict the binding of TFs. Chromatin accessibility is controlled by regulatory proteins known as pioneer factors, which have the ability to bind nucleosome rich regions and make them accessible to other TFs. Figure adapted from Shlyueva et al. (2014).

Although multiple regulatory elements are involved in controlling the gene expression, enhancers play one of the most important role in transcriptional regulation because of their ability to activate and enhance the transcription, and drive cell-type specific expression patterns. From here onwards, this chapter is focused on enhancers and their associated mechanisms. The following sections provide greater details about the function of enhancers, approaches to predict enhancer regions, and discuss how disruptions in these regions can lead to diseases.

1.2 Enhancers

1.2.1 History

The first enhancer region was identified in 1981, as a 72 bp sequence segment of the simian virus 40 (SV40) genome (Banerji et al., 1981). This sequence could increase the transcription levels of a reporter gene (a β-globin gene) by two hundred times in HeLa cells (a cell line derived from cervical cancer cells). Surprisingly, it was observed that this enhancer sequence could influence the transcription of the reporter gene from different locations in either orientation; it could increase the transcription levels of the reporter gene when placed near the promoter, or several kbs away, upstream or downstream of the gene (Banerji et al., 1981). Subsequent to this discovery, many studies confirmed these observations and identified more enhancers in animal viruses (Hansen and Sharp, 1983; Schirm et al., 1985; Spandidos and Wilkie, 1983; Villiers et al., 1982) and metazoan genomes (Banerji et al., 1983).

One of the first enhancers in the mammalian genome was described in the immunoglobulin heavy chain (IGH) locus. The function of this IGH associated enhancer

sequence was observed to be cell dependent; the enhancer was functionally active in myeloma cells (derived from cancerous plasma cells), but not in HeLa cells (Banerji et al., 1983; Davidson et al., 1986). A similar observation was made with enhancers associated with β-globin gene, where enhancer function was not only cell-type specific, but also stage-specific in the developmental process (Antoniou et al., 1988; Kollias et al., 1987; Trudel and Costantini, 1987). Initially, enhancers were considered to be structural elements associated with chromatin organisation, but this assumption soon changed as evidence about cell-type specific activity of enhancers emerged. It was hypothesised that the cell-dependent activity of enhancers was a result of the presence or absence of cell-type specific TFs binding to the enhancers. Indeed, around end of the 1980s, further studies using *in vivo* and *in vitro* techniques characterised the binding sites of several TFs and demonstrated that enhancer function is dependent on specific TF binding activity (Lee et al., 1987; Maniatis et al., 1987).

It is now 37 years since the first discovery of enhancer in the SV40 genome, the core definition of enhancers has not changed significantly (Pennacchio et al., 2013; Tippens et al., 2018). The completion of the human genome project and the falling cost of high-throughput sequencing technologies have facilitated the identification of enhancers on a genome-wide scale. In the last decade, many research groups and large-scale projects such as the ENCODE (ENCODE Project Consortium, 2012) and the Roadmap Epigenome Project (Bernstein et al., 2010), have identified up to a million potential enhancers in a plethora of human cell-types and tissues by employing high-throughput sequencing technologies. Such projects have not only produced a wealth of useful epigenetic data, but also increased the interests of researchers in epigenetics. As a result, enhancers are currently considered as the most important class of functional elements in the non-coding part of the genome.

1.2.2 Enhancer function

Enhancers are responsible for increasing the transcription rate of their target genes. Moreover, most of the mechanisms that drive cell-type specific gene expression at different stages of the developmental process are believed to be regulated by enhancers (Zlotorynski, 2018). Enhancers achieve this by recruiting cell-type specific TFs and co-factors. An important property of enhancers is that their function is independent of their orientation and distance with respect to their target gene, hence they are capable of exhibiting their effects on a gene located several hundred kbs or Mbs away through chromosome looping (Gondor and Ohlsson, 2009) (explained later in section 1.2.4). For successful transcription, a gene requires the PIC to be assembled at its promoter region, which will initiate the transcription and overpower pol II pausing resulting

in transcription elongation (Fig. 1.5). However, the level of basal transcription by promoters is often low. The enhancers come in contact with the promoters via looping and increase the rate of transcription by increasing the number of factors involved in the process. Most important factors amongst these include the Mediator complex, which is a co-activator complex binding to other TFs and pol II (Kagey et al., 2010); cohesin, which stabilises and sometimes even drive cell-type specific enhancer-promoter communication bridge (Kagey et al., 2010); and factors important for paused pol II release and elongation such as *BRD4* (Liu et al., 2013). Additionally, enhancers can exhibit their effect in an additive or partially redundant manner on the overall transcription of their target genes. This characteristic is evident from *in vivo* assays (such as reporter assays) where amalgamating multiple enhancer sequences often display transcription levels equivalent to their combined effect (Arnone and Davidson, 1997).

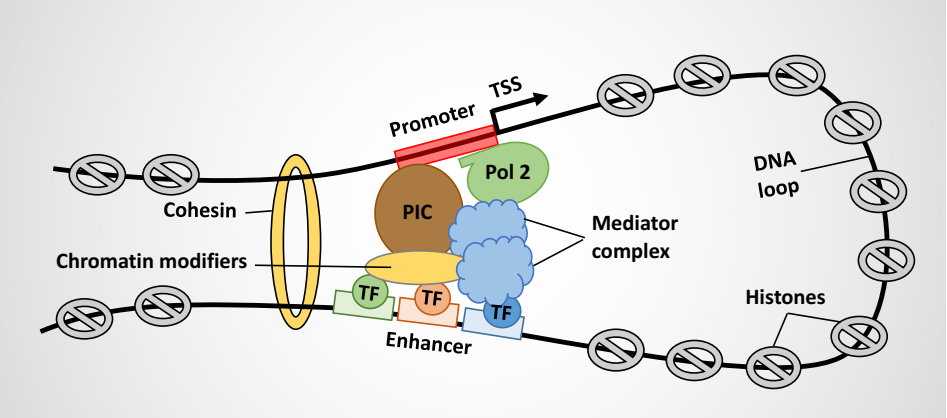


Fig. 1.5 Enhancer function. Enhancers once bound by the required TFs become active and up-regulate the expression of their target genes. Enhancer function involves the formation of a DNA loop in order to bring enhancers into close proximity with the promoters of their target genes. This DNA loop is believed to be mediated by cohesin and the Mediator complex. Figure adapted from Heinz et al. (2015).

1.2.3 Enhancer states and their associated chromatin marks

Enhancer activation starts with binding of TFs and remodelling of the nucleosome within the enhancer region. The binding of TFs is followed, and in some scenarios assisted by, the binding of other co-regulators such as p300 and CREB-binding protein (Wang et al., 2009); pol II; chromatin remodellers such as the *BRG1* complex (Morris et al., 2014); and the Mediator complex (Kagey et al., 2010). These events lead to the modifications of histone tails (such as methylation and acetylation) present in the immediate enhancer related nucleosome. Interestingly, it has been observed that some of these histone modification signatures are specifically associated with enhancers (Table

1.1). For example, enhancer sites are enriched in histone H3 lysine4 monomethylation (H3K4me1) or histone H3 lysine4 dimethylation (H3K4me2), and lack histone H3 lysine4 trimethylation (H3K4me3), while active enhancer sites have the addition of histone H3 lysine27 acetylation (H3K27ac) (Creyghton et al., 2010; Heintzman et al., 2007). Additionally, enhancers marked by the repressive mark histone H3 lysine27 trimethylation (H3K27me3) are considered to be poised (Zentner et al., 2011).

Table 1.1 Histone tail modifications and their presumed biological associations.

Histone mark	Functional association
H3K4me1	Enhancers and distal elements, and regions downstream of transcription start sites
H3K4me2	*Cis*-regulatory regions and promoters of transcriptionally active genes
H3K4me3	Promoters and transcription start sites
H3K27ac	Active regulatory elements
H3K9ac	Active regulatory elements with preference for promoters
H3K27me3	Elements repressed by polycomb proteins
H3K79me2	Transcription, with preference to 5' end of genes
H3K9me3	Repressive heterochromatin and repetitive elements
H3K36me3	Actively transcribed portions of genes and chromatin regions.

As an example, the enhancer activity and epigenetic features associated with T-cell acute lymphocytic leukaemia 1 (*TAL1*) in multiple human cell lines are shown in Fig. 1.6. Enhancers B and C, representing an enhancer stretch of 19 kb downstream and 3.8 kb upstream of *TAL1* TSS respectively, activate *TAL1* transcription in human umbilical vein endothelial cells (HUVECs) (Gottgens et al., 2004), while in K562 erythroid cells, enhancer A (51 kb downstream of *TAL1* TSS) and enhancer C are responsible for doing the same (Delabesse et al., 2005). The enrichment of H3K4me2 and H3K27ac is consistent with the cell-type specific activity of the enhancers. Note that the cell lines which do not express *TAL1*, such as human embryonic stem cells (hESCs), normal human epidermal keratinocyte (NHEK) and normal human lung fibroblast (NHLF), lack histone modifications associated with active enhancers, while displaying high enrichment of the repressive mark H3K27me3.

As opposed to enhancers, active promoter regions have an enrichment of H3K4me3 and H3K27ac, and a depletion of H3K4me1 (Heintzman et al., 2007; Kim et al., 2005). Other signatures such as enrichment of histone H3 lysine79 trimethylation (H3K79me3) and pol II, have been associated with active enhancers which regulate the genes involved in the developmental process (Bonn et al., 2012). It is important to note that although not all the regions with these histone modification signatures are functionally active enhancers, but the majority of active enhancers have been observed to have these characteristics. For this reason, such epigenetic marks have been utilised by researchers to annotate and differentiate between the various enhancer states (Ernst and Kellis, 2012). The enhancers have been broadly categorised into the following four states (Fig. 1.7):

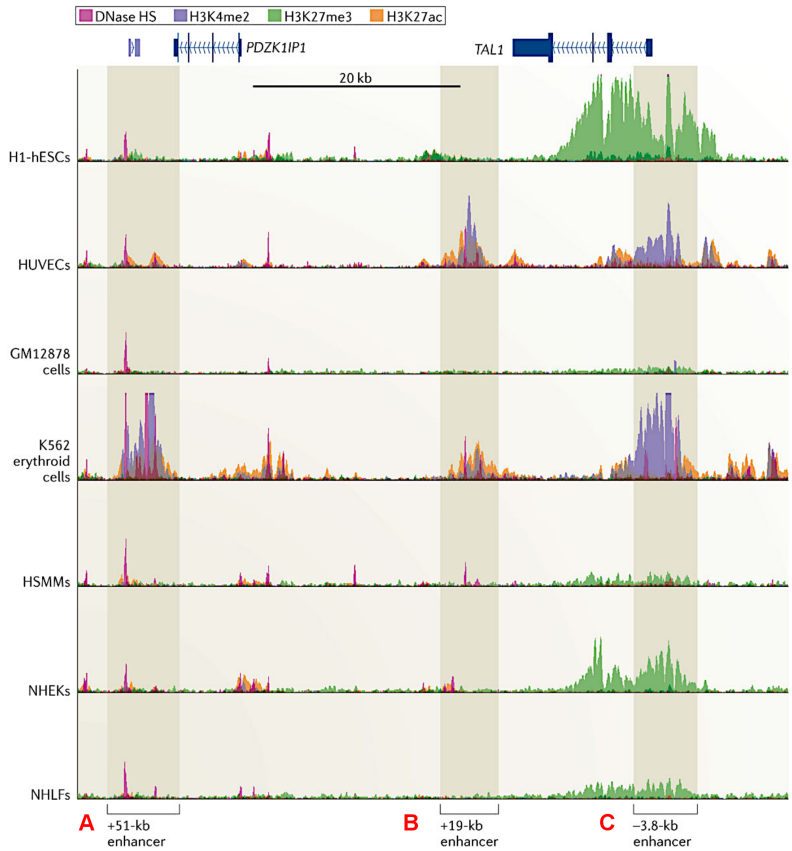


Fig. 1.6 Cell-type specific activity of histone modification signatures. Genomic view of ∼60 kb region around the *TAL1* gene, displaying levels of H3K4me2, H3K27me3, H3K27ac and DNase HS (DNaseI hypersensitive sites which represent open chromatin regions) in seven human cell lines. hESC, human embryonic stem cell; HUVECs, human umbilical vein endothelial cells; HSMM, human skeletal muscle myoblast; NHEK, normal human epidermal keratinocyte; NHLF, normal human lung fibroblast. Figure taken from Heinz et al. (2015).

1. **Inactive**: An inactive enhancer is covered with closed or compact chromatin which makes the region inaccessible to protein binding, hence they lack TF and co-regulator activity (Fig. 1.7A).

2. **Primed**: A primed enhancer has open chromatin, and therefore, this region may be bound by TFs. However, these TFs depend on other additional factors such as external stimulus, other TFs and co-regulators, to activate the enhancer function. Histones near primed enhancers may have H3K4me1 modifications, but are depleted of H3K27ac (Fig. 1.7B).

3. **Poised**: A poised enhancer is essentially a primed enhancer additionally characterised by repressive chromatin marks such as H3K27me3, and may have a small open chromatin region (Fig. 1.7C).

4. **Active**: Active enhancers have an accessible chromatin and are bound by all the essential TFs required for enhancer activation. These elements are marked by a high enrichment of H3K27ac and H3K4me1(Fig. 1.7D).

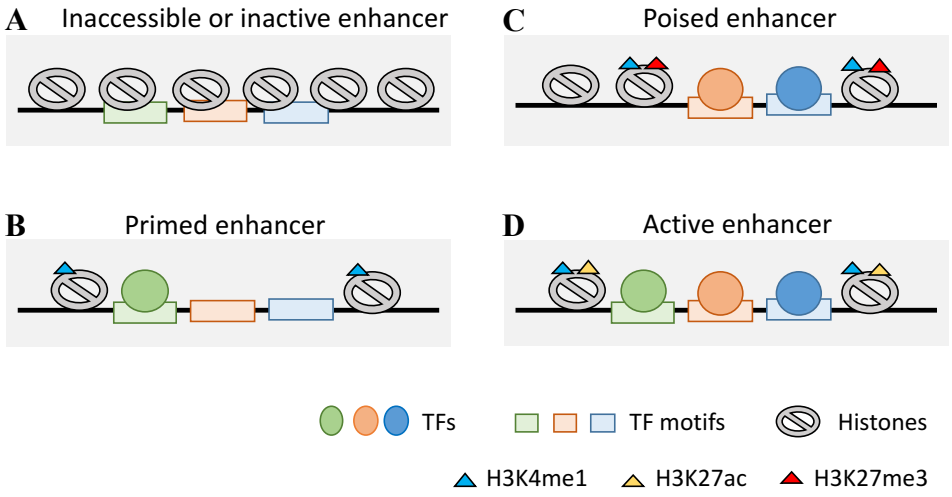


Fig. 1.7 Various enhancer states. (A) Inactive enhancers usually consists of high density of nucleosomes which make them inaccessible to TFs. (B) Enhancers which are not yet active and depend on other signals to get activated may be enriched with H3K4me1. (C) Poised enhancers may be marked by H3K4me1 and repressive mark H3K27me3. (D) Active enhancers are bound by TFs and have high enrichment of H3K4me1 and H3K27ac. Figure adapted from Shlyueva et al. (2014).

1.2.4 Enhancer-promoter interaction

Soon after the discovery of enhancers, it was proposed that enhancers interact with the promoters of their target genes even from hundreds of kbs away, which raised the question of how do enhancers find and interact with promoters from such long distances. Over the years, two major models have been proposed to explain this mechanism (Fig. 1.8). First, a facilitated-tracking model, which states that once an enhancer is bound by activator proteins, it moves along the DNA in the direction of promoter, thus forming a loop which increases in size until the enhancer-bound complex comes into contact with the promoter (Blackwood and Kadonaga, 1998; Wang et al., 2005). Second, a looping model which suggests that there occurs a more direct contact between the genomic location of the enhancer and the promoter within the nucleus, by looping out of the DNA (Gondor and Ohlsson, 2009). This chromosome looping is believed to be mediated by cohesins (Hadjur et al., 2009; Hou et al., 2010) and other TFs like *GATA1* (Vakoc et al., 2005), Mediator (Kagey et al., 2010) and *CTCF* (Hou et al., 2010). Enhancer RNAs (eRNAs), which are a by-product of enhancers being transcribed by pol II, are also believed to be physically engaged in the formation of such loops (Hsieh et al., 2014).

Since there could exist large distances between enhancer and promoter elements, it is unlikely that the chromatin structure between the two elements spanning over such large distance is directly involved in the enhancer-promoter interaction. Hence,

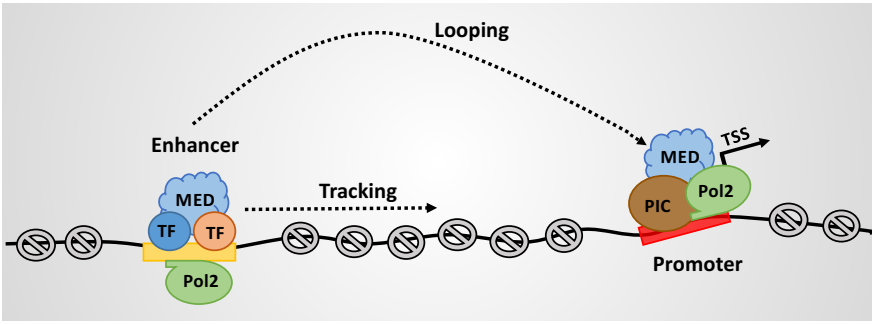


Fig. 1.8 Mechanisms of enhancer-promoter communication. The tracking model proposes that enhancer-bound TF complex moves towards the promoter of their target gene without leaving the enhancer sequence. The looping model proposes that the enhancer-bound TF complex loops out of the intervening DNA sequence to interact with the promoter of its target gene. Figure adapted from Vernimmen and Bickmore (2015).

the facilitated-tracking model is likely to be applicable for close enhancer-promoter interactions (< 10 kb) in comparison to long-range enhancer-promoter communications, where the chromosomal looping model is more appropriate. Indeed, the few studies which provide evidence for a facilitated-tracking model involved a relatively small distance between the enhancer and promoter (Hatzis and Talianidis, 2002; Wang et al., 2005). Presently, the facilitated-tracking model is only supported by a small number of studies and requires further testing.

In contrast to the facilitated-tracking model, the looping model has been widely tested experimentally using fluorescence *in situ* hybridisation and chromosome confirmation capture (3C) type techniques, which have provided evidence for the long-range enhancer-promoter interactions through chromatin looping. The 3C type techniques includes 3C (Lieberman-Aiden et al., 2009) and its variants such as circular chromosome confirmation capture (4C) (Zhao et al., 2006), chromosome confirmation capture carbon copy (5C) (Dostie et al., 2006) and Hi-C (Belton et al., 2012). These 3C based assays are able to capture regions in the genome which are physically close to each other, independent of their proximity on the linear DNA. For instance, the β-globin locus was amongst the first group of genes where DNA looping interactions were identified between β-globin promoters and a LCR situated 25 kb upstream of them (Carter et al., 2002). Another 3C variant method called chromatin interaction analysis with paired-end tag sequencing (ChIA-PET) (Fullwood et al., 2009) has been widely used, which integrates the chromatin immunoprecipitation (ChIP) technique (Das et al., 2004) with 3C based assays to capture chromosomal contacts involved with protein factors important in transcription, such as the pol II. Such studies have shown that gene promoters have frequent long-range interactions with multiple enhancers (which are often cell-type specific) to form active chromatin hubs (Laat and Grosveld, 2003). Genome-wide analysis of chromatin interactions using 3C based techniques (termed Hi-

C) have also identified that genome is partitioned into active and inactive compartments (Lieberman-Aiden et al., 2009). These compartments have been further sub-divided into domains at the sub-megabase level called topologically associated domains (TADs). TADs represent regions of the genome having high chromatin interactions and have been found to be conserved amongst humans and mouse (Dixon et al., 2012).

1.2.5 Enhancer target genes

The enhancer-promoter interaction maps from chromosome capture studies are also being utilised to identify potential genes regulated by enhancers. The identification of the target gene regulated by a particular enhancer has been a challenging task in regulatory genomics. Generally in the past, the nearest annotated gene to the enhancer on the linear genome has been assumed to be its potential target gene. But, even the genomic regions separated by several Mbs on the linear genome might be located close to each other in the 3D organisation of the genome, therefore, the gene nearest to the enhancer may not be its correct target gene in many cases. Indeed, the data from 3C based chromosome interactions suggest that enhancers often skip their nearest annotated gene and regulate distantly located genes. However, the results from such studies are not consistent and report that the proportion of enhancers that regulate their nearest gene ranges between 7% up to >80%. Some of the 3C based studies and their findings are shown in Table 1.2. Contrastingly, a study based on co-expression patterns to predict regulatory interactions show that 88% of functional enhancers in *Drosophila* regulate their nearest gene, with 8% of the remaining genes interacting with their second nearest gene (Kvon et al., 2014). Additionally, they found that up to 80% of intragenic enhancers in *Drosophila* regulate the gene within which they are located, which suggests that enhancers more than often regulate the gene they overlap. Overall, the functional relevance of the genome-wide enhancer-gene interactions obtained from 3C based studies has been difficult to completely validate due to the limitations described below.

Firstly, the resolution of capturing the chromosome interactions is often low (5 kb - 1 Mb) as it depends on the sequencing depth and restriction enzyme digestion sites. Although, recent methods such as *in situ* Hi-C have increased the resolution up to 1 kb by performing very deep sequencing (over 25 billion reads) (Rao et al., 2014), it is expensive and computationally intensive to produce and analyse such large amount of reads to achieve a high resolution. Hence, due to the low resolution in the majority of the studies, it is possible that the chromosome confirmation is not able to reliably capture the contacts between adjacent genomic elements (<5 kb), where many enhancers could be located (Yao et al., 2015). For instance, using the capture Hi-C method (CHi-C) (Mifsud et al., 2015) which specifically involves enrichment of reads around the promoters,

Table 1.2 Examples of 3C based studies and their observations.

Data	Organism	Cell line/tissue	Finding	Reference
5C	Human	GM12878, K562, HeLa-S3	7% of all interactions are with the nearest TSS; 27% of distal elements interact with their nearest gene, or 47% if only expressed genes are used in the analysis	Sanyal et al., 2012
ChIA-PET	Human	MCF7, K562, HeLa, HCT116, NB4	60% of enhancers interact with their nearest gene	Li et al., 2012
ChIA-PET	Mouse	mESCs	83% of SEs and 87% of TEs interact with their proximal active gene	Dowen et al., 2014
CHi-C	Human	GM12878, CD43+	66% of all interactions are with the nearest promoter	Mifsud et al., 2015

a higher frequency of interactions (66%) was identified to interact with the nearest promoter (Table 1.2). Besides, many studies have shown that enhancers often do not regulate and skip the genes not expressed in the cell-type under study (Mifsud et al., 2015; Sanyal et al., 2012), suggesting that the interactions between enhancer and their target genes may depend on the genomic or epigenomic circumstances.

Secondly, the 3C based methods were initially developed to capture the physical contacts in the genome, which may not necessarily correspond to functional regulatory interactions (reviewed in Laat and Duboule (2013)). Indeed, many studies have found frequent inter- and intra-chromosomal interactions within genomic regions with no chromatin activity (Lieberman-Aiden et al., 2009; Sanyal et al., 2012). For instance, a 5C study in fetal lung cells identified 1 million interactions, out of which only $\sim$6% were detected between an annotated promoter and a distal region (Jin et al., 2013). It is also believed that many of these interactions which are not related to gene expression may arise from random collision in the nucleus or within the TADs, or they may be involved in maintaining the nuclear structure, hence are stable but not functional (Dekker et al., 2013). It is also possible that some enhancers interact with the promoters of their target genes by a different mechanism rather than looping. Therefore, the 3C based studies may not provide a complete set of chromatin interactions between potential enhancers and their target genes (Yao et al., 2015).

Although chromosome confirmation studies have provided immense insights into the spatial organisation of the genome, the rules to infer enhancer-promoter interactions are not yet clear and require further investigation using *in vivo* genetic approaches. Since there is a paucity of chromosome confirmation data in different cell-types, many genome-wide studies associate potential enhancers to their nearest genes, in order to obtain enough enhancer-promoter interactions to provide them with useful biological insights (Hnisz et al., 2013; Hnisz et al., 2015; Loven et al., 2013; Whyte et al., 2013).

1.3 Current methods to identify enhancer regions

The identification and characterisation of enhancers has been of great interest amongst researchers since their discovery. Especially, in the last decade, studying enhancers have been of prime focus in the area of genomics, revealing their importance not only in regulating cell-type and stage-specific gene expression, but also in disease causation (discussed later in this chapter). However, identification of enhancers has been a challenging task. One of the reasons for this is that identifying enhancers and their activity solely from DNA sequences is unreliable. Moreover, it is difficult to identify essential regions within enhancers and infer their functional consequence in the event of disruption. But, with the advancement in next-generation sequencing technologies during the last decade, it has now become possible to reliably capture enhancer related chromatin features at a genome-wide scale. In addition, techniques to modify genomic DNA *in vivo* allow researchers to experimentally investigate enhancer activity and its functional consequences. In the following section, I describe some of the common approaches which are used to predict enhancer regions and how they take advantage of the known characteristics of enhancers to identify them.

1.3.1 Predictions using motifs and conservation

As described earlier, enhancers are known to contain binding sites for TFs, which are often observed to be conserved between closely related species. Therefore, computationally searching for TF binding motifs within a genomic sequence can help to predict enhancer regions. Two approaches have been used to achieve this: (1) by identifying genomic sequences enriched for TF binding motifs; or (2) by identifying genomic sequences highly conserved between species. Both these methods depend on prior knowledge of motif sequences bound by TFs. There are many computational and experimental methods to identify the motif sequence of a TF. The computational approaches mainly involve looking for short sequences either over-represented or with high evolutionary sequence conservation amongst potential regulatory regions such as a group of functionally related sequences. For instance, examining the promoter regions upstream of co-regulated or co-expressed genes for enriched and conserved sequences is a common technique to identify motifs (Roth et al., 1998).

However, the computational approaches described above have limitations. Since motifs are short DNA sequences generally 6-12 bp in length, they could even match to random genomic sequences which can produce a high rate of false-positives. In reality, only a small fraction of these matches would be bound by the TF *in vivo* (Wasserman and Sandelin, 2004). Additionally, TF binding is often tissue-specific and relies heavily

on interactions with other co-factors and proteins (Yanez-Cuna et al., 2012), which suggests that only a small fraction of computationally identified motif matches would have functional relevance. Likewise, motif matches with high evolutionary sequence conservation do not guarantee that they would be bound by the TF in the tissue of interest, or that they would be functionally active. Previous studies have shown that enhancers can be functional and weakly conserved between the species (Blow et al., 2010; Meireles-Filho and Stark, 2009), therefore, methods to predict enhancers solely based on conservation may capture only a portion of the total enhancers active genome-wide.

1.3.2 Predictions using TF binding

Compared to computational scanning of motifs to predict transcription factor binding sites (TFBSs), a more accurate approach is to identify the *in vivo* binding sites of a TF. Many experimental approaches have been used to identify genome-wide *in vivo* TFBSs. These methods include chromatin immunoprecipitation followed by sequencing (ChIP-seq) (Johnson et al., 2007), and its variants such as ChIP-exo (Rhee and Pugh, 2011) and DNA adenine methyltransferase identification (DamID) (Steensel and Henikoff, 2000). Of these methods, ChIP-seq has been the most commonly used technique. ChIP-seq involves the chemical crosslinking of TFs to their *in vivo* binding sites and enrichment of these DNA-protein complexes using antibodies specific to the TFs (Fig. 1.9). This is followed by deep sequencing of the retrieved DNA fragments and computational analysis to identify genomic regions bound by the TF. A major limitation of ChIP based methods is that they require a highly specific antibody for the TF of interest, and antibodies for many TFs have not been discovered yet. An additional limitation is the cost, since only one TF can be profiled in a single ChIP-seq experiment.

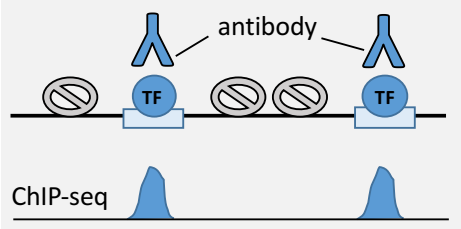


Fig. 1.9 Identification of transcription factor binding *in vivo*. ChIP-seq uses antibodies specific to the TFs to identify the location of their genome-wide DNA binding sites. Figure adapted from Shlyueva et al. (2014).

An analysis of a single ChIP-seq experiment detects on average thousands of *in vivo* binding sites, most frequently occurring within promoters, introns and intergenic regions (Spitz and Furlong, 2012). From ChIP-seq data of hundreds of TFs, we know

that TF binding pattern actively changes during development, indicating that TFs act in a time and cell-type specific manner (Yanez-Cuna et al., 2012). Generally, a ChIP-seq experiment is able to detect the majority of functional binding sites of a TF (Zeitlinger et al., 2007). However, not all the binding sites identified by ChIP-seq are functional and therefore, would not correlate with functional enhancers (Fisher et al., 2012). These observations have revealed that a TF can be bound to the DNA without influencing the expression of any gene. Although the exact reason for this is unclear, it is believed that these non-functional binding sites may be dependent on cooperative binding and still require binding of other TFs to activate their function. Moreover, TFs tend to have a general affinity towards the DNA and may bind to regions with open chromatin outside their functional network at low levels or for a very short period of time.

1.3.3 Predictions using chromatin accessibility

Since active enhancers lack nucleosomes and contain a loosely packed chromatin, identification of such regions on the DNA can help in predicting enhancers and other regulatory elements. Techniques such as DNaseI hypersensitive sites sequencing (DNase-seq) (Boyle et al., 2008) and micrococcal nuclease digestion followed by sequencing (MNase-seq) (Yuan et al., 2005) have been mostly used for this purpose. These techniques use enzymes such as DNaseI or micrococcal nuclease respectively, which have the property to cleave the DNA at nucleosome depleted regions (Fig. 1.10). Therefore, these techniques are able to capture the complete accessible or open chromatin landscape of a cell or tissue, and predict enhancers independent of TF information. This is particularly important as DNase-seq or MNase-seq could be used to predict enhancers even in those cells for which critical lineage-specific TFs are not yet discovered. In addition to DNase-seq and MNase-seq, other techniques such as formaldehyde-assisted identification of regulator elements followed by sequencing (FAIRE-seq) (Giresi et al., 2007) and assay for transposase accessible chromatin followed by sequencing (ATAC-seq) (Buenrostro et al., 2015) have been developed recently to quantify chromatin accessibility. FAIRE-seq is based on the fact that formaldehyde crosslinking is stronger at nucleosome-occupied regions, compared to nucleosome-free regions. Whereas ATAC-seq uses a Tn5 transposase to place sequencing adapters into open chromatin regions of the DNA (Buenrostro et al., 2015).

The TFBSs identified using ChIP-seq have been observed to be significantly correlated with open chromatin regions detected using DNase-seq or FAIRE-seq (Kaplan et al., 2011; Pique-Regi et al., 2011). Although, the majority of open chromatin regions overlap with TFBSs which may represent enhancer regions, not all open chromatin regions correspond to active enhancers. Other regulatory proteins such as insulators,

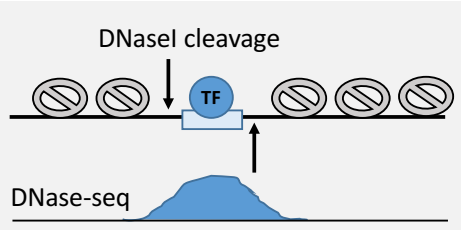


Fig. 1.10 Identification of genome-wide open chromatin regions. DNase-seq uses the enzyme DNaseI which has the property to cleave the DNA regions depleted of nucleosomes. These regions are referred to as DNaseI hypersensitive sites (DHSs). Figure adapted from Shlyueva et al. (2014).

also bind at regions of open chromatin. For instance, the CCCTC-binding factor (also know as *CTCF*) with enhancer-blocking insulator properties is detected within open chromatin regions, but does not function as an enhancer (Xi et al., 2007). Moreover, the core promoter region near the TSSs have open chromatin as TFs and co-regulators bind there (Xi et al., 2007). Lastly, open chromatin regions may also be inactive due to the binding of repressive TFs, a phenomenon commonly observed during development (Gray and Levine, 1996). For such reasons, in order to predict enhancers, the open chromatin information is often combined with other characteristics of enhancers such as histone modifications.

1.3.4 Predictions using histone modifications

Researchers have widely used histone modifications to predict enhancer activity, taking advantage of the fact that distinct histone tail modifications occur near enhancers and other regulatory elements (described above in section 1.2.3). Like chromatin accessibility, predicting enhancers using histone modifications is independent of TF information and can be applied to any cell-type or tissue (Fig. 1.11). For this reason, ChIP-seq of histone marks has been the most commonly used method adopted by large international genomic projects (Bernstein et al., 2010; ENCODE Project Consortium, 2012; Ernst et al., 2011; Kharchenko et al., 2010; Shen et al., 2012) to map genome-wide enhancers in different organisms and these predictions have shown to correlate well with enhancer activity assays (Arnold et al., 2013; Heintzman et al., 2007). Algorithms like ChromHMM (Ernst and Kellis, 2012) have enabled us to better annotate the different enhancer states by combining data from multiple histone marks into easy interpretable annotations. For example, by modelling together eight histone marks and the *CTCF* binding profile, Ernst et al. (2011) were able to categorise the genome into 15 chromatin states corresponding to active and inactive enhancers and promoters. Such systematic genome-wide annotation of enhancers using histone modifications have even discovered

novel enhancer states, such as enhancers enriched with both active and repressive histone marks (H3K4me1 and H3K27me3 respectively) termed as bivalent enhancers (Bernstein et al., 2006; Shu et al., 2011).

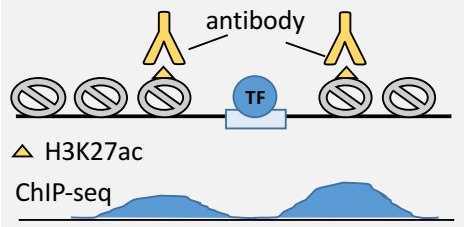


Fig. 1.11 Identification of genome-wide chromatin marks. Histones flanking enhancer regions contain specific modifications such as H3K27ac. ChIP-seq using antibodies specific to these histone marks is used to detect the genome-wide locations of histone modifications. Figure adapted from Shlyueva et al. (2014).

Despite histone modifications being widely used to identify potential enhancers, it is not yet known what proportion of the complete enhancer landscape can be captured using this method. It has been observed that the predictions from known histone modifications and their combinations may not perfectly correlate with enhancer activity (Arnold et al., 2013; Bonn et al., 2012). For instance, a study in *D.melanogaster* detected ∼40% of mesodermal enhancers lacked H3K27ac signal (Bonn et al., 2012), however, these numbers may vary in different organisms. In addition, novel histone acetylation marks such as H3K64ac and H3K122ac, have been recently discovered to correlate with promoter and enhancer activity (Pradeepa et al., 2016). Though data for these novel histone modifications is not abundant at the moment, such additional marks may improve enhancer prediction in the near future.

1.3.5 Predictions using enhancer-promoter interactions

It has now been established that enhancers come in contact with promoters of the gene they regulate, therefore, identification of these interactions or the characteristics associated with them can also help in enhancer prediction. This concept has been utilised by two methods. The first method is based on the fact that co-factors help mediate the enhancer-promoter interactions, hence, identification of genomic binding sites of such co-factors can predict enhancers. Although, the complete mechanism of the 3D organisation of the genome still remains to be determined, scaffold proteins like Mediator and cohesin are known to stabilise the chromosome looping (Kagey et al., 2010). Therefore, ChIP-seq profiles of Mediator and cohesin have been used to predict enhancers (Whyte et al., 2013) (Fig. 1.12). The second method is the 3C based assays (describer earlier in section 1.2.4) which directly quantifies both intra- and

inter-chromosomal physical contacts between enhancers and the core promoters of their target genes (Fig. 1.12). Although, chromosome confirmation capture techniques have been used to identify enhancers and predict their regulatory interactions with the target genes, there are some limitations associated with this methodology which are described earlier in section 1.2.5.

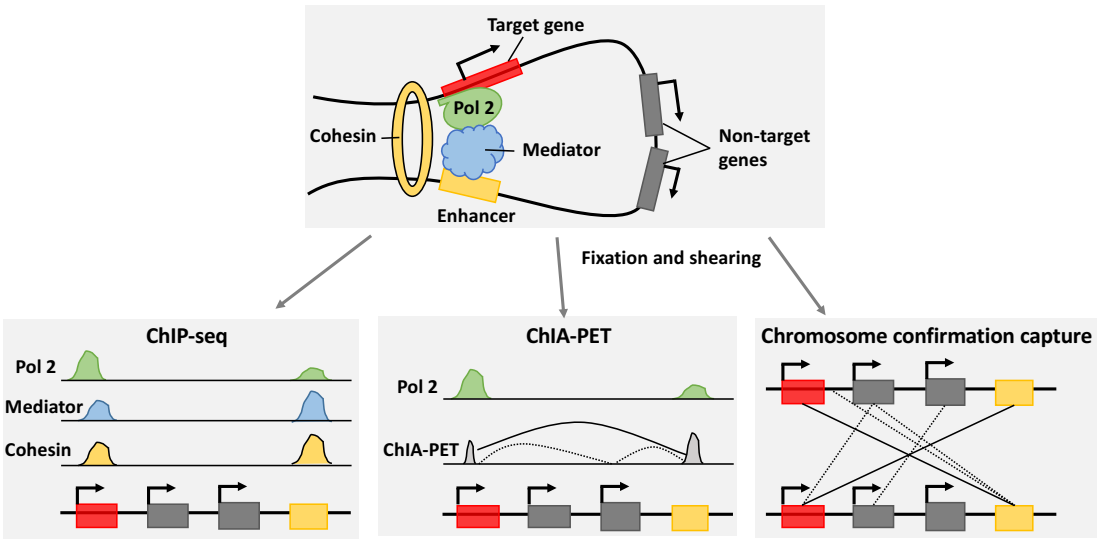


Fig. 1.12 Identification of enhancer-promoter interactions. Enhancers are brought into close proximity of the promoters of their target genes through chromosome looping, which is mediated by Mediator complex and cohesin. ChIP-seq can be used to identify the contact points between Mediator complex and cohesin. ChIA-PET and 3C based methods involve formaldehyde crosslinking of spatial contacts, shearing of linear DNA, fragmentation, ligation and deep sequencing. ChIA-PET additionally involves a ChIP phase for the enrichment of contacts which involve a particular protein such as the pol II. 3C based methods detect all the spatial and physical contacts within a defined genomic region. Both ChIA-PET and 3C based methods can identify pairwise interactions between the contact points (shown as lines), some of which may correspond to regulatory interactions (solid lines) and some may not (dotted lines). Figure adapted from Shlyueva et al. (2014).

1.4 Super-enhancers

A single enhancer is enough to activate the transcription of its target gene. But some genes are located in regions near a high density of enhancers. Such genes have been often observed to be expressed in a tissue- or cell-type specific manner. A well characterised example of such a region is the enhancer rich LCR associated with tissue-specific expression of globin genes in erythroid cells (Levings and Bungert, 2002). These dense clusters of active enhancers have been termed as super-enhancers (SEs) and recognised as a new class of regulatory element.

The term 'super-enhancer' was first used in 2004, corresponding to a 30 bp enhancer sequence which could activate the transcription of immediate-early gene (*ie-1*) both *in vitro* and *in vivo* (Chen et al., 2004). More recently, genome-wide SEs have been identified and characterised using ChIP-seq, to detect regions densely bound by key TFs. In comparison to regular enhancers (referred to as typical-enhancers (TEs)), SEs span relatively large genomic regions (up to tens of kb) and have significantly higher enrichment of chromatin regulators, TFs and co-factors such as the Mediator complex, *p300*, *Brd4* and pol II (Hnisz et al., 2013).

1.4.1 Super-enhancer identification

SEs were first distinguished from TEs in mouse ESCs (mESCs) using *Med1* binding signal by Whyte et al. (2013). This classification was achieved in three steps: (1) binding sites of key TFs associated with pluripotency in mESCs (such as *Sox2*, *Oct4*, *Nanog*) were identified using ChIP-seq and considered as potential enhancer regions; (2) enhancer regions within a distance of 12.5 kb were stitched together (Fig. 1.13A); and (3) *Med1* binding profile was used to identify stitched enhancers with significantly high enrichment of *Med1* signal and were defined as SEs (Fig. 1.13B). In order to discover SEs in differentiated cells, key lineage-specific TFs have been commonly used instead of pluripotency associated TFs. For example, the binding occupancy of the TF *PU.1* was used by Whyte et al. (2013) to identify potential enhancer regions in pro-B cells and further categorise them into SEs. This methodology to detect SEs was later implemented by Whyte et al. (2013) in a program called ROSE (Rank Ordering of Super-Enhancers), which has since become the standard method to do such analysis and has been applied to an assortment of cells and tissues (Table 1.3). However, the marks used to identify potential enhancers and segregate SEs have been variable. Overall, Mediator binding and H3K27ac chromatin mark has been most commonly used to segregate SEs from TEs.

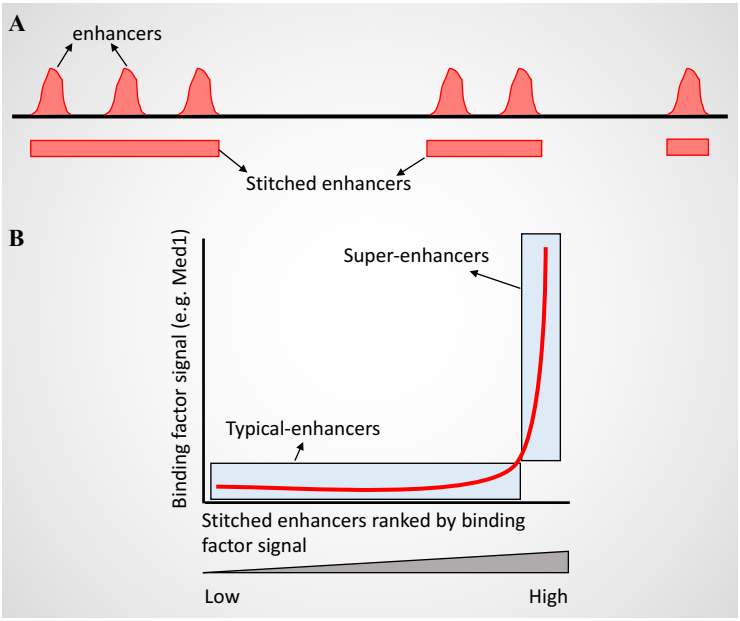


Fig. 1.13 Identification of super-enhancers. (A) The enhancer regions identified in any cell-type or tissue within a distance of 12.5 kb are stitched together into cohesive units. (B) The stitched enhancer units are then ranked by a binding factor enrichment signal (such as *Med1* or H3K27ac) and a threshold of its inflection point is calculated. The stitched enhancer units with a binding factor signal higher than the estimated threshold are defined as SEs. Figure adapted from Pott and Lieb (2015).

1.4.2 Properties of super-enhancers

SEs have been associated with their nearest genes based on the observation that enhancer-promoter looping interactions frequently occur within 50 kb (Gondor and Ohlsson, 2009; Ong and Corces, 2011; Sanyal et al., 2012; Spitz and Furlong, 2012). Systematic mapping of SEs across diverse tissues and cell lines in human and mouse has shown that SEs regulate key genes that define the cell identity, and drive high expression of their target genes compared to TEs (Adam et al., 2015; Fang et al., 2015; Hnisz et al., 2013; Huang et al., 2016; Loven et al., 2013; Ohba et al., 2015; Pelish et al., 2015; Shin et al., 2016; Siersbæk et al., 2014; Vahedi et al., 2015; Whyte et al., 2013). For example, important genes involved in ESCs pluripotency such as *Oct4*, *Sox2* and *Nanog*, are found near SEs (Whyte et al., 2013). In another study, *Sox2* associated SE was identified to be essential for 90% of *Sox2* expression (Li et al., 2014). Genes associated with SEs have also been observed to be expressed in a cell-type specific manner (Loven et al., 2013). Furthermore, compared to genes associated with TEs, SE associated genes appear to be more sensitive to perturbations. For example, shRNA knockdown of *Med12* mostly affected the expression of SE associated genes, however, the degree of impact on the expression levels of TE associated genes was much lower.

Table 1.3 Summary of previous studies which involved identification of super-enhancers. Table adapted from Niederriter et al. (2015).

Enhancer identification	Factor to distinguish SE and TE	Tissue/Cell	Organism	Reference
1. Oct4, Sox2, Nanog 2. Stitch together	Med1	mESC	Mouse	Whyte et al., 2013
		mESC	Mouse	Hnisz et al., 2015
1. Med1 2. Stitch together		MM1.S cell line SCLC cells Glioblastoma cells	Human	Loven et al., 2013
		Multiple AML cell lines	Human	Dawson et al., 2014
1. H3K4me1/DHS 2. Stitch together		Erythroid cells	Mouse	Hay et al., 2016
1. PU.1 2. Stitch together	PU.1	Pro-B cells	Mouse	Whyte et al., 2013
1. MyoD 2. Stitch together	MyoD	Myotubes	Mouse	Whyte et al., 2013
1. T-bet 2. Stitch together	T-bet	T-helper cells	Mouse	Whyte et al., 2013
1. C/EBPA 2. Stitch together	C/EBPA	Macrophages	Mouse	Whyte et al., 2013
1. EBNA2 2. Stitch together	EBNA2	EBV-transformed lymphoblastic cells	Human	Zhou et al., 2015
1. H3K27ac 2. Stitch together	H3K27ac	86 tissues/cell lines	Human	Hnisz et al., 2013 Suzuki et al., 2017
		Colorectal cancer cells ER+ breast cancer cells	Human	Hnisz et al., 2015
		Jurkat cells	Human	Dawson et al., 2014
		MOLM-1 cells	Human	Groschel et al., 2014
		Neuroblastoma cells	Human	Chipumuro et al., 2014
		T-cells	Mouse	Vahedi et al., 2015
		EBV-transformed lymphoblastic cells	Human	Zhou et al., 2015
		ESCs, Pro-B cells, Th cells, myotubes, macrophages	Mouse	Suzuki et al., 2017
	Not described	Striatum	Mouse	Achour et al., 2015
1. BRD4 2. Stitch together	BRD4	B cell lymphoma	Human	Chapuy et al., 2013
		Activated endothelial cells	Human	Brown et al., 2014
1. STAT5/H3K27ac 2. Stitch together	Med1/H3K27ac/GR	Mammary tissue	Mouse	Shin et al., 2016

1.4.3 Controversy over super-enhancer structure and function

Since the first characterisation of SEs by Whyte et al. (2013), many studies have identified SEs in a wide range of tissues and cells. However, for most tissues *Med1* data is not available and lineage-specific TFs are often undiscovered, therefore different co-factors (e.g. p300) and chromatin marks (e.g. H3K27ac) have been utilised for SE characterisation (Table 1.3), hence leading to inconsistent methodologies. Many researchers also remain sceptical over the method of SE identification, arguing that the clustering of enhancers on the linear genome into SEs and TEs solely based on chromatin marks enrichment is arbitrary and lacks functional relevance (Gray and Levine, 1996; Pott and Lieb, 2015). This is because such a clustering could group together enhancers which regulate different genes and conversely, it could group enhancers which regulate the same gene into different clusters.

Another aspect of SEs which has been questioned is its novelty; it is believed that SEs show previously known characteristics of enhancers and overlap with previously defined regulatory elements. For instance, SEs in K562 cells overlap with the well characterised LCR of the human β-globin locus (Hnisz et al., 2013). Similar to SEs, clusters of potential regulatory elements termed as 'clusters of open regulatory elements' (COREs), have been previously described using open chromatin regions and linked to tissue-specific TFs (Gaulton et al., 2010; Song et al., 2011). Another previously described enhancer category similar to SE is a 'stretch enhancer' (Parker et al., 2013). First characterised around the same time as SEs, stretch enhancers were defined as enhancer regions (non-stitched) greater than or equal to 3 kb in length. However, a comparison between stretch enhancers and SEs in the same cell-types showed that stretch enhancers are significantly higher in number and SEs make up only a small proportion of stretch enhancers ($\sim$3%), suggesting SEs to be a subset of stretch enhancers (Niederriter et al., 2015). Based on these observations, some researchers believe that SEs may be counterparts of previously defined regulatory elements, with differences in number and genomic position arising due to the different criteria used to define such elements.

With respect to the SE function, it is not yet clear whether individual enhancer elements within a SE work in an additive (total contribution towards target gene expression equal to the sum of the strength of its individual elements), synergistic (total contribution towards target gene expression greater than the sum of the strength of its individual elements), redundant (total contribution towards target gene expression less than the sum of the strength of its individual elements) or a more complicated manner. Recent advances in genome editing techniques such as CRISPR-Cas9 (short for clustered regularly interspaced short palindromic repeats and CRISPR-associated protein 9) mediated deletions (Shalem et al., 2014), have facilitated researchers to explore the impact of individual SE elements on their target gene transcription by deleting them. For instance, *in vivo* deletion of individual enhancers within the α-globin SE showed that two out of five individual enhancers significantly contributed to α-globin expression in an independent and additive manner (Hay et al., 2016). Whereas, a similar study in *Wap* associated SE showed its expression to be partially dependent on each of the individual SE elements, which the authors referred to as a functional hierarchy within the SE (Shin et al., 2016). Furthermore, Moorthy et al. (2017) observed that each individual enhancer within enhancer clusters contributes to the expression of their associated genes (*Dppa5a*, *Ooep*) and hence, have partially redundant function.

Conversely, Hnisz et al. (2015) investigated the function of individual and combinations of enhancer elements within *Pou5f1* SE using reporter assays, which showed that the individual enhancer elements within *Pou5f1* SE neither have additive nor synergistic effect when present in a single copy, instead they exhibit a complicated influence on each

others activity. Recently, another study performed similar experiments in SEs associated with three micro-RNAs (*miR-290-295* in mESCs, *miR-1* in myotubes and *miR-148a* in Pro-B cells) by generating cell lines depleted of individual enhancer elements within the SEs, which again showed a cooperative effect amongst individual SE elements, rather than an additive or redundant effect (Suzuki et al., 2017). Interestingly, genome-wide chromatin interaction data from ChIA-PET suggests that individual enhancer elements within SEs have more frequent interactions compared to elements within TEs (Dowen et al., 2014). Clearly, similar studies are required to learn and understand about the influence of individual SE elements on each other and on their target gene expression, especially on a genome-wide scale.

1.5 Mis-regulation of enhancer function in disease

1.5.1 Early examples of enhancer malfunction in disease

Given that the enhancers play a key role in transcriptional regulation, it is of no surprise that apart from the changes in the protein-coding portion of the genome, any kind of disruption either in the enhancer regions, or in TFs that directly interact with enhancers could attribute to diseases. One of the earliest instances of the involvement of a regulatory region in disease was identified in the blood disorder thalassaemia. A DNA translocation disrupting the LCR associated with the β-globin locus was identified to be responsible for β-thalassaemia (Kioussis et al., 1983). This LCR located approximately 25 kb upstream of β-globin locus, is mostly composed of enhancers, which are together responsible for driving high expression of the β-globin genes. Later, using 3C based methods, these enhancers were shown to have long-range interactions with the promoter of β-globin genes (Tolhuis et al., 2002). In another example, mutations in a limb specific enhancer, situated 1 Mb upstream of sonic hedgehog gene (*SHH*) leads to polydactyly (abnormal limb development) (Lettice et al., 2002). These examples show that mutations or other disruptions within enhancers can lead to enhancer loss of function.

1.5.2 Enhancers in cancer and other diseases

Early studies exploring disease-causing mutations were performed mostly on a single gene, and investigated simple phenotypic traits in Mendelian diseases such as cystic fibrosis. However, these methods could not be directly applied to diseases with a complex genetic nature such as diabetes and obesity. But, over the past decade, powerful statistical methods have been developed to associate genetic variants to complex diseases and traits. Genome-wide association studies (GWASs) have been a key resource for this

purpose analysing millions of genome-wide genetic variants (usually single nucleotide polymorphisms (SNPs)) to test their association with both Mendelian and complex diseases. GWASs are based on linkage disequilibrium (LD, measured as a squared correlation r^2); an event of two SNP alleles co-occurring close to each other is not random and exists as a result of natural selection and recombinant rate (Slatkin, 2008). A typical GWAS involves two groups of individuals; one which are affected by a particular disease trait under study, and the other which are not affected by that trait (control set). Both groups are then genotyped for SNPs and allelic frequencies of their SNPs are compared between the two groups using rigorous statistical tests, which identifies SNPs with strong association to the disease trait under investigation (typical thresholds used: p-value $= 5 \times 10^{-8}$; $r^2 > 0.5$).

To date, GWASs have reported causal genetic variants for hundreds of Mendelian and complex diseases, which include both common diseases and traits that are risk factors for diseases (Fig. 1.14). In addition to disease traits, GWASs have also identified genetic variants associated with physical traits such as height (Cousminer et al., 2013), hair colour and skin pigmentation (Han et al., 2008). Although, early studies focused on characterising SNPs within protein-coding sequence of the genes, over 90% of the disease-associated SNPs (DA-SNPs) from GWASs occur within the non-coding regions of the genome (Hindorff et al., 2009; Manolio, 2010). This observation made it apparent that SNPs within the non-coding regions contribute to disease causation as well. With the availability of genome-wide regulatory maps in human, 64% of these non-coding SNPs have been found to be within enhancers (H3K27ac enriched regions) (Hnisz et al., 2013). Similarly, $\sim$76% of the non-coding SNPs from GWASs have been found to occur within DHSs or in high linkage disequilibrium with a SNP within DHS (Maurano et al., 2012). Furthermore, the DA-SNPs more than often occur in regulatory regions active in the cell-types linked to the disease pathology. For instance, SNPs associated with Alzheimer's disease have been identified in brain-specific enhancers, while those associated with coronary heart defects have been identified in heart-specific regulatory regions (Hnisz et al., 2013; Maurano et al., 2012). This tendency of DA-SNPs to occur in cell-type specific regulatory regions has explained how sequence variants, which occur in all cells, lead to certain cell- or tissue-specific disease traits.

Although GWASs can detect genetic variants significantly associated with a disease trait, it does not provide any functional information about the mechanism underlying this association. Therefore, further characterisation of GWAS hits is required to understand how these genetic variants in enhancer regions lead to gene aberrations and diseases. Recent studies have started to functionally characterise the causal genetic variants in regulatory regions to understand the pathways they disrupt, which could be utilised to develop more efficient disease therapies. Generally, the non-coding variants within

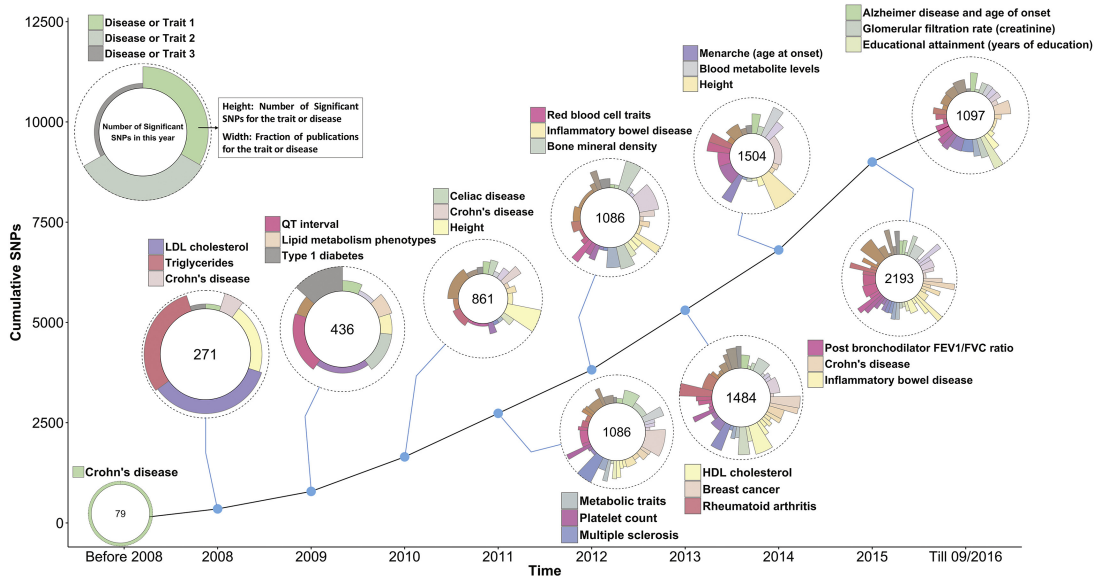


Fig. 1.14 Timeline of SNPs discovered by GWASs. For each year, the top three diseases and traits with the highest number of SNPs are labelled. Figure taken from Visscher et al. (2017).

enhancer regions can disrupt the binding sites of sequence-specific TFs, which could either completely remove a binding site from that location, or create a novel binding site for other TFs, ultimately affecting the expression of their associated target gene. For instance, intronic SNPs associated with fetal haemoglobin expression levels within the gene *BCL11A* have been detected to alter the binding sites of *GATA1* and *TAL1*, which disrupts the binding patters of these TFs resulting in reduced expression of *BCL11A* and fetal haemoglobin (Bauer et al., 2013). Similarly, many studies have now identified and characterised SNPs in enhancers to be associated with common diseases (Farh et al., 2015; Gjoneska et al., 2015; Hnisz et al., 2013; Maurano et al., 2012; Pasquali et al., 2014). In addition to enhancer regions, many diseases have been associated with mutations causing mis-regulation in TFs, co-factors and chromatin regulators, which directly or indirectly interact with enhancers (reviewed in Herz (2016)). Altogether, mutations in regulatory regions have been identified to contribute to several disease areas such as cancer, neurological disorders, autoimmunity, diabetes and many more (examples shown in Table 1.4).

Table 1.4 Examples of enhancers linked to human disease.

Disease	Implicated gene	Type	Reference
Associated with enhancer			
Aniridia	*PAX6*	Rearrangement	Kleinjan et al., 2001
X-linked deafness type 3	*POU3F4*	Deletion, rearrangement	De Kok et al., 1995, De Kok et al., 1996
Van Buchem disease	*SOST*	Deletion	Loots et al., 2005
Campomelic dysplasia	*SOX9*	Rearrangement	Pfeifer et al., 1999
Celft lip	*IRF6*	Point mutation	Rahimov et al., 2008
Multiple cancer types	*MYC*	Point mutation	Jia et al., 2009, Pomerantz et al., 2009, Ahmadiyeh et al., 2010, Sotelo et al., 2010
Prostate cancer	*RFX6*	Point mutation	Huang et al., 2014
Parkinson's disease	*SNCA*	Point mutation	Soldner et al., 2016
Type 1 diabetes	*FOXP3*	Point mutation	Bassuny ET AL., 2003
Associated with co-factors and chromatin-regulators			
Uterine leiomyomas; prostate cancer	*MED12*	Point mutation	Makinen et al., 2011, Barbieri et al., 2014
Multiple cancer types	*ARID1A*	Point mutation	Hargreaves and Crabtree, 2011
Non-Hodgkin lymphoma, B-cell lymphoma	*EZH2*	Point mutation	Morin et al., 2011, Pasqualucci et al., 2011
Rubinstein-Taybi Syndrome; multiple cancer types	*CBP, EP300*	Point mutation	Roelfsema et al., 2005, Lawrence et al., 2014
Type 2 diabetes	*PPARGC1A*	Point mutation	Ling et al., 2008

1.5.3 Super-enhancers in human diseases

Super-enhancers in complex disease

SEs in human cell-types have been identified to frequently harbour DA-SNPs compared to TEs. Furthermore, the DA-SNPs tend to occur in disease-relevant cell-types (Hnisz et al., 2013). For instance, in an analysis of ∼5000 SNPs from GWASs, 19% (5/27) of the SNPs associated with Alzheimer's disease occurred in brain SEs; 19% (13/67) of the SNPs associated with type 1 diabetes occurred in primary T-helper cell SEs; and 33% (22/67) of the SNPs associated with systemic lupus erythematosus occurred in B-cell SEs (Hnisz et al., 2013). In all the above mentioned diseases, the enrichment of their associated SNPs was significantly higher in SEs compared to TEs. A similar pattern was observed for other diseases such as multiple sclerosis and rheumatoid arthritis; and traits like white blood cell distribution and fasting insulin level.

Many other independent studies found DA-SNPs to occur in SEs linked with disease-relevant genes. For example, a GWAS performed on vitiligo (an autoimmune skin disease) patients detected three DA-SNPs close to each other in a SE between *HLA-DRB1* and *HLA-DQA1* genes (Cavalli et al., 2016). Likewise, inhibition of a SE region in primary T-cells specifically associated with juvenile idiopathic arthritis (an autoimmune disease) resulted in reduced expression of the disease-related genes (Peeters et al., 2015). In addition to SEs, SNPs associated with type 2 diabetes (T2D) have also been detected to overlap stretch enhancers annotated in pancreatic islet cells (Parker et al., 2013). Moreover, impaired SE function can be caused by altered H3K27ac levels and pol II binding, as detected in a mouse model of Huntington's disease causing aberrant regulation of striatal neuronal genes (Achour et al., 2015; Le Gras et al., 2017). These examples provide evidence that genetic or epigenetic changes in SEs contribute to complex diseases via mis-regulated transcription of their associated genes.

Super-enhancers in cancer

Soon after the discovery of SEs in 2013, SEs were characterised in multiple myeloma tumour cells (Loven et al., 2013). Loven et al. (2013) observed that SEs are associated with many oncogenes, including *MYC*, a gene which is commonly expressed at high levels in many cancers. Similarly, Chapuy et al. (2013) observed SEs in diffuse large B-cell lymphoma to be associated with many previously known oncogenes. Interestingly, when the SE landscape was compared between tumour and healthy cells, it was observed that oncogenes acquire SEs in tumour cells (Hnisz et al., 2013). These *de-novo* SEs are believed to be acquired as a result of small indels, chromosomal translocation, focal

amplification, over-expression of oncogenic TFs or somatic mutations (examples shown in Table 1.5). Furthermore, eRNAs transcribed from SEs have also been shown to be associated with tumorigenesis (Jiao et al., 2018; Li et al., 2013; Liang et al., 2016; Teppo et al., 2016). Overall, SEs have been observed to be involved in the regulation of genes which play a role in cancer progression.

Table 1.5 Super-enhancers in cancer.

Cancer type	Implicated gene	Mechanism/ observation	Reference
T-ALL	*TAL1*	Small indels introduce novel binding sites for *MYB* resulting in a *de novo* SE	Mansour et al., 2014
Multiple myeloma		Translocation of 3' IgH SEs; translocation of breakpoints resulting in a *de novo* SE	Hnisz et al., 2013; Walker et al., 2014
Lung cancer, SCLC		Tandem repeats within SE of *MYC*; focal amplification of enhancers near *MYC*	Hnisz et al., 2013; Iwakawa et al., 2013
AML, lung adenocarcinoma, endometrial carcinoma	*MYC*	Focal amplification of a large SE downstream of *MYC*	Shi et al., 2013 Zhang et al., 2016
T-ALL		Overexpression of *TAL1* TF within SE associated with *MYC*	Hnisz et al., 2013
AML	*EVI1*	Translocation of *GATA2* SE	Groschel et al., 2014
Adenoid cystic carcenoma	*MYB*	Translocation of SEs increase expression of *MYB*	Drier et al., 2016
Neuroblastoma	*LMO1*	Somatic mutation in intron causing differential *GATA* binding	Oldridge et al., 2015

T-ALL: T-cell acute lymphoblastic leukaemia; SCLC: small cell lung cancer; AML: acute myeloid leukaemia.

Targeting super-enhancers for cancer therapeutics

A common characteristic of most cancer cells is that they have high oncogenic transcriptional activity compared to healthy cells, which help them to grow at a faster rate (Lin et al., 2012). Therefore, suppressing transcriptional activity of specific oncogenic targets is believed to be an effective clinical therapeutic. Since SEs regulate genes involved in tumour progression, attempts are underway to target SE activity to reduce oncogenic transcriptional activity. In order to achieve this, researchers are utilising small molecules to inhibit specific essential components of SEs to disrupt their influence on oncogenes (summarised in Table 1.6).

SEs in myeloma cells were observed to exhibit high enrichment of *MED1* and *BRD4* binding (Loven et al., 2013). *BRD4* is a member of bromodomain and extra-terminal (BET) family proteins, which bind to the Mediator complex and interact with pol II, hence, are a critical element for SE associated transcription (Hnisz et al., 2013; Zeng and Zhou, 2002). Many BET bromodomain inhibitors (such as JQ1 and iBET) are currently under study to target *BRD4* activity within SEs. The first study to show the effect of BET inhibitor on SE activity was conducted in myeloma cells using the BET inhibitor JQ1 (Loven et al., 2013). This study demonstrated that treating myeloma cells with JQ1 significantly reduced *BRD4* binding (up to 97%) in SEs, leading to decreased *MED1* binding, increased pol II pausing, and ultimately decreased expression of genes such as the oncogene *MYC*. Subsequent to this finding, similar studies were performed in other cancer types and diseases to analyse the effect of JQ1. For example, JQ1 has been observed to reduce the expression of genes associated with juvenile idiopathic arthritis in T-cells (Peeters et al., 2015), and slow down the tumour growth in adenoid cystic carcinoma (Drier et al., 2016). Apart from JQ1, other BET inhibitors like iBET, have also been found to be effective in reducing tumour growth in acute myeloid leukaemia (Pelish et al., 2015) and neuroblastoma (Wyce et al., 2013).

In addition to BET inhibitors, researchers have also used small molecules to inhibit cyclin-dependent kinases (CDKs) in SEs, which control the pol II initiation and elongation (Malumbres, 2014). For instance, the THZ1 inhibitor can effectively inhibit *CDK7* (Kwiatkowski et al., 2014), and its treatment has shown to selectively decrease expression of *MYC* family and other oncogenes in neuroblastoma cells (Chipumuro et al., 2014), small cell lung cancer cells (Christensen et al., 2014), esophageal squamous cell carcinoma (Jiang et al., 2017) and T-cell leukaemia cells (Wong et al., 2017). Moreover, techniques like CRISPR-Cas9 facilitate gene therapy strategies in diseases associated with SEs. For instance, Mansour et al. (2014) used CRISPR-Cas9 to delete the somatic mutations responsible for the formation of a SE, and abolish its effect on *TAL1* expression. Although safety and efficiency of such approaches require further extensive research in animal models, they provide new opportunities to develop effective gene therapies. Overall, these examples demonstrate that targeting SEs can be used to selectively inhibit expression of oncogenes, and such approaches have application in disease diagnosis and developing cancer related therapeutics.

Table 1.6 Therapeutic targeting of super-enhancers in cancer. Table taken from Sengupta and George (2017).

Cancer type	Inhibitor	Effect on SE-driven transcription	Effect of SE inhibition on tumour biology	Reference
DLBCL	JQ1 *(BRD4)*	Downregulation of SE-driven oncogenic and lineage-specific transcriptional circuits.	Decreased lymphoma infiltration in the bone marrow and improved overall survival	Chapuy et al., 2013
AML	JQ1 (*BRD4*)	Eviction of *BRD4* and Mediator from select SE regions causing decreased expression of associated genes that are *MYB* targets and important for leukemogenesis	Impaired proliferation and triggering differentiation of leukemic blasts	Bhagwat et al., 2016
Oncogenic *Nras* expression in mouse liver	iBET *(BRD4)*	Reduced expression of genes involved in SASP that are driven by SEs.	Decreased clearance of oncogenic senescent cells.	Tasdemir et al., 2016
T-ALL, MYCN-amplified NB, SCLC, TNBC	THZ1 *(CDK7)*	Downregulation of SE-associated and tumour addictive and lineage specific gene expression, *MYCN*-driven transcriptional amplification	Decreased tumour volumes, growth and increased survival	Chipumuro et al., 2014, Kwiatkowski et al., 2014
AML	Cortistatin A *(CDK8/19)*	Upregulation of SE-associated genes linked to tumour suppression and lineage specification.	Reduction in disease progression, leukemic burden, and tumour volume, improved overall survival.	Pelish et al., 2015
T-ALL	THZ531 *(CDK12/13)*	Downregulation of DNA damage response and SE-associated genes	Apoptosis	Zhang.T et al., 2016
Ewing sarcoma	LEE011 *(CDK4/6)*	Downregulation of SE-associated ES dependency genes CyclinD1/*CDK4*	Cytostasis and delayed growth	Kennedy et al., 2015

DLBCL: diffuse large B-cell lymphoma; AML: acute myeloid leukaemia; SASP: senescence-etory phenotype; T-ALL: T-cell acute lymphoblastic leukaemia; NB: neuroblastoma; SCLC: small cell lung cancer; ES: Ewing sarcoma.

1.6 The Mouse as a model organism

For many decades, genes associated with disease traits in human have been investigated in model organisms such as the mouse. Studying human diseases in model organisms has been vital for understanding the biological function of genes. For this purpose, animal models have either been generated by disrupting genes orthologous or equivalent to disease-causing genes in human, or that display phenotypic features similar to the disease condition in humans. Such models have proven to be successful in enhancing our knowledge of Mendelian disorders and penetrant mutations. Animal studies are particularly useful as they: (1) allow to take repeated phenotypic measurements within an environmentally and genetically controlled background; (2) could be used for tissues not accessible from human patients; and (3) could be used for developing and testing new drugs. Of all the animal models, the mouse has been the prime mammalian model to study human diseases because of their high genetic and physiological similarities to humans (Nguyen and Xu, 2008). Mouse models of human diseases have provided novel critical insights into disease mechanisms, as these models display very similar phenotypic characteristics to the pathological condition in humans (Schofield et al., 2012). The following sections describe how mouse models have helped in the functional annotation of mammalian genes, and in validating and characterising enhancer sequences.

1.6.1 Mammalian phenotypes

The pathological or disease characteristics in mouse models are commonly described as phenotypes, which could be defined as an observable trait showing deviation from normal morphology, physiology or behaviour. In order to avoid ambiguity amongst these phenotype terms and to allow their efficient computational analysis, they are organised into a formal hierarchal structure of controlled vocabulary called ontologies. For example, Gene Ontology (GO) (Ashburner et al., 2000) describes the biological processes, molecular functions and cellular locations of the gene products. Similarly, the most widely used ontology for mouse phenotypes is the Mammalian Phenotype (MP) Ontology (Smith et al., 2005). The MP terms describe abnormal phenotypes and other phenotypic measures in an animal, which are deviant from the control population. It is important to note that the MP terms are not equivalent to any specific disease, but a group of MP terms may describe the characteristic features of a disease. The hierarchy based structure of the MP ontology allows mouse phenotype databases to be queried for mutations and alleles associated with a specific phenotype, and also enables the researchers to identify clusters of genes related to similar phenotype terms, which may represent genes in the same functional pathway.

1.6.2 Large scale phenotyping projects

Mouse models have been most commonly generated by: (1) N-ethyl-N-nitrosourea (ENU) induced mutations, which involves using the chemical mutagen ENU to induce random point mutations; (2) targeted mutations to alter the gene function (also known as knockins); or (3) targeted mutations to completely eliminate the gene function (also known as knockouts or null mutation). The ENU mutagenesis projects are based on the phenotype-driven approach where the transgenic mice carrying the ENU-induced mutations are first phenotypically screened to identify a clinical phenotype of interest, and then other strategies such as positional mapping or genome sequencing are employed to identify the gene harbouring the causal ENU mutation. Large scale ENU mutagenesis projects in the past have generated novel mouse models for human diseases across several phenotype areas (Brown and Nolan, 1998; Hrabe de Angelis et al., 2000; Masuya et al., 2005; Nolan et al., 2000; Potter et al., 2016; Thaung et al., 2002), which have played a key role in discovering disease-associated genes. On the other hand, targeted mutations are based on the genotype-driven approach where a researcher with some prior knowledge about a gene, alters its gene structure via targeted mutations, and investigates its role in the phenotype of interest. While previous studies involving targeted mutations have been helpful to understand the gene function, they have mainly focused on screening mouse models for a specific phenotype domain of interest. However, a gene may be responsible for performing different functions depending on where it is expressed in the body, or its time of expression during the life span of an organism. This phenomenon is commonly known as pleiotropy.

To identify such pleiotropic functions of genes, many large scale gene knockout projects such as the EUMODIC (The European Mouse Disease Clinic) (Ayadi et al., 2012), the SANGER-MGP (The Wellcome Trust Sanger Institute Mouse Genetics Project) (Ayadi et al., 2012) and the IMPC (International Mouse Phenotype Consortium) (Brown and Moore, 2012a), have been established in the last decade to extensively phenotype mouse knockout lines in order to discover gene-phenotype associations which have been previously undetected. EUMODIC and SANGER-MGP initiated in the late 2000s together have phenotyped approximately 800 mutant mouse lines (Ayadi et al., 2012). The IMPC project started in September 2011, with the aim to produce extensive phenotyping data for a knockout of every protein-coding gene in the mouse genome ($\sim$20,000 genes) (Brown and Moore, 2012b). The IMPC project involves generating a knockout mouse line for each protein-coding gene and then screening them through a systematic pipeline of phenotype tests to capture all the phenotypes associated with the gene in study. The pipeline of phenotype tests includes standardised protocols covering a wide range of biological systems tracked via the database IMPReSS (International Mouse Phenotyping Resource of Standardised Screens; http://www.mousephenotype.

org/impress). Phenotypic measurements are compared between the mutant mouse lines and controls, and statistically significant phenotypes associated with a gene are identified and are annotated using the MP ontology (Angelis et al., 2015). To date, IMPC have phenotyped over 6,000 knockout mouse lines and over 5,000 genes, producing ∼58,000 phenotype annotations (data release 9.2). The ultimate goal of IMPC is to build an encyclopaedia of gene function for all the protein-coding genes in the mammalian genome, and make both the mice and the data publicly available to the research community which would provide a platform for further investigation.

1.6.3 Functional testing of enhancers in the mouse

The majority of the mouse models in the past involved studying protein-coding genes, but recently, mouse models have also been used to understand how enhancer sequences function *in vivo*. Previously, reporter gene constructs have been used to characterise the activity of endogenous elements such as promoters, that can drive expression (Kothary et al., 1989). These constructs when microinjected into fertilised mouse eggs, merge with the genome and the transgenic mice can then be screened for the activity of the element. A similar approach has been used for testing the activity of enhancers, with reporter gene constructs containing the candidate enhancer sequence (to be tested) upstream of the minimal promoter. The enhancer if active, drives the expression of the reporter gene which corresponds to the endogenous enhancer activity. Such enhancer assays in mouse have been fundamental to identify and validate enhancers, and to understand their functional properties. Large-scale enhancer screens have been performed to characterise the activity of candidate enhancers in whole mouse embryos (at E-11.5 days) by whole mount staining, and the results of such studies to date are stored in the VISTA enhancer browser (Visel et al., 2007). This database containing whole mount images of enhancer activity has helped researchers to identify cell-type specific markers, and select functional enhancers for further characterisation (Gordon et al., 2014; Sanchez-Castro et al., 2013).

Many enhancers are dependent on their genomic or chromatin context (Arnold et al., 2013; Kvon et al., 2014) because of which some enhancer sequences with endogenous activity may show no/weak activity in transgenic assays, due to the enhancer being outside of their native chromatin context. It has also been shown that some enhancers selectively work with only specific types of core promoters (Butler and Kadonaga, 2001; Zabidi et al., 2015). These limitations associated with the traditional enhancer-reporter assays can be overcome by using transgenes based on large bacterial (BACs), yeast (YACs) or P1-derived artificial chromosomes (PACs) (Giraldo and Montoliu, 2001), as they drive gene and reporter expression from native promoters, hence closely recapitu-

lating endogenous gene expression. BACs have been most commonly used for model organisms like mice. Collectively, enhancer assay studies in mouse have shown that enhancers are capable of driving highly specific and dynamic gene expression profiles, which are critical for the mammalian development. Furthermore, these studies have demonstrated that conserved non-coding DNA sequences often display enhancer activity *in vivo* (Pennacchio et al., 2006), and that enhancers involved in the developmental process (particularly the ones active in forebrain) are highly conserved across species (Nord et al., 2013).

In the last decade, the enhancer assays in mouse have also been utilised to functionally validate enhancer predictions from massively parallel sequencing technologies such as ChIP-seq and DNase-seq (Cotney et al., 2012; Nord et al., 2013; Visel et al., 2009). Even enhancer sequences from the human genome have been tested in mouse enhancer assays. For instance, 66% of the enhancers identified in the human heart tissue successfully drove reporter gene expression when integrated with the mouse genome (May et al., 2011). Another advantage of *in vivo* assays is that they allow the testing of enhancer activity in the relevant disease-associated tissues, whereas *in vitro* models may fail to detect the relevant enhancer activity. This has been useful to investigate the effects of disease-associated genetic variants. One recent example includes the obesity associated variants identified within an enhancer in the *FTO* gene, which were demonstrated to be connected with *IRX3* expression in mouse white adipose tissue (Smemo et al., 2014). Furthermore, mouse models generated using genome editing techniques have been extremely insightful to understand the phenotypic effects of enhancer deletions in their native chromatin context. Several studies have used transgenic mice with targeted deletion of enhancers or their components, to investigate its effect on gene expression at the whole organism level (Canver et al., 2015; Cunningham et al., 2018; Dickel et al., 2018a; Hay et al., 2016; Shin et al., 2016; Sur et al., 2012). Overall, such studies in transgenic mice have been critical to understand that the impact and contributions of individual enhancer elements on gene expression is complex and difficult to predict at the whole organism level.

1.6.4 MRC Harwell Institute

The MRC Harwell Institute (MRCHI) specialises in the use of mouse models to study the relationships between genes and diseases. The MRCHI conducts large scale phenotyping screens namely, the IMPC, and an ENU ageing mutagenesis screen to study the genetics of ageing. The research programs at the MRCHI focuses on different areas which can be broadly divided into lifetime studies, translational studies, data analysis and dissemination. Although these research programs have traditionally con-

centrated on investigating functional consequences of coding-variants in mouse models, recently there has been more emphasis on studying regulatory elements in mutant mice. The above research and mutagenesis programs are supported by the data analysis and dissemination groups: statistical genomics and biocomputing, which also carry out independent research in their respective fields. My DPhil was within the biocomputing group and I also worked closely with the neurobehavioral and metabolic genetic groups at the MRCHI during my DPhil.

1.7 Aims of the thesis

Since the completion of the human genome sequence, protein-coding genes have been extensively studied to understand their function and involvement in diseases. However, less is known about the function of non-coding regions in the genome and their functional implication in diseases. The ENCODE project has begun to address this gap and has shown that a large portion of the genome is involved in the regulation of genes, directly or indirectly, and predicted up to a million potential enhancer regions in humans (Thurman et al., 2012). Other studies have found DA-SNPs to frequently occur within the enhancer regions of disease-relevant tissues and cell-types (Hnisz et al., 2013). Therefore, it important to understand how transcription is controlled by regulatory elements, especially enhancers and thus, is one of the main emerging challenges in genomics today.

The overall aim of this thesis is to investigate the effect of regulatory regions, especially enhancers, on mouse models of human diseases. To address this, my goal is to:

1. Systematically identify potential active enhancers and promoters in the mouse genome using publicly available data and compare their activity across multiple tissues to identify tissue-specific regulatory elements.

2. Identify super-enhancers in the mouse genome and explore their functional characteristics compared to typical-enhancers.

3. Investigate how different enhancer architectures influence gene expression and tissue-specificity of their associated target genes.

4. Characterise the functional association of different enhancers in disease aetiology by analysing mouse phenotypes and diseases linked with the enhancer-associated genes.

5. Develop novel methods to integrate publicly available omics data with functional data at the MRCHI, in order to study transcriptional regulatory mechanisms in mouse models currently under investigation at the MRCHI.

Chapter 2

Klf14 transcriptional networks in human and mouse

In this chapter, I describe the investigation of *Klf14* associated regulatory pathways in the human and mouse genomes. This work was carried out in collaboration with the McCarthy lab at the Oxford Centre for Diabetes, Endocrinology and Metabolism (OCDEM), the Small lab at King's College London and the Cox lab at the MRCHI. Some results described in this chapter (section 2.2.1 and 2.2.2) have been published in the following article:

Small, K. S., M. Todorčević, M. Civelek, J. S. El-Sayed Moustafa, X. Wang, M. S. Simon, J. Fernandez-Tajes, A. Mahajan, M. Horikoshi, A. Hugill, C.A. Glastonbury, L. Quaye, M. J. Neville, **S. Sethi**, et al. (2018). "Regulatory variants at KLF14 influence type 2 diabetes risk via a female-specific effect on adipocyte size and body composition". In: *Nature Genetics* 50.4, pp. 572-580. ISSN: 1546-1718. DOI: 10.1038/s41588-018-0088-x.

2.1 Introduction

A primary objective of human genetics is to identify genetic variants which cause diseases and phenotypic traits in the human population, and functionally understand how these variants lead to a disease state. During the last decade, GWAS has proven to be an important tool of human genetics and has produced a wealth of genomic regions which show strong statistical association with a wide range of human diseases and phenotypic traits (Visscher et al., 2017). Amongst these human diseases, metabolic traits have been widely analysed in GWASs especially type 2 diabetes (T2D). T2D is a condition where the pancreatic beta cells fail to produce the right amount of insulin

which results in increased glucose levels in the body (Taylor, 2013). The worldwide prevalence of diabetes has almost doubled in the last three decades (from ∼4% in 1980 to ∼8% in 2014) (NCD-Risk-Factor-Collaboration, 2016), with ∼90% of people estimated to have T2D. Both genetic (with an estimation of 25%-69% heritability) (Almgren et al., 2011; Poulsen et al., 1999; Willemsen et al., 2015) and environmental factors such as sedentary lifestyle, high calorific foods and gut flora (Friedman, 2009; Moreno-Indias et al., 2014), are believed to be the cause of T2D. This expanding prevalence has revealed a requirement for a deeper understanding of the T2D mechanisms, a better understanding about its aetiology, leading to better treatment or possible prevention in the future.

Despite a pool of genetic variants potentially associated with T2D risk, the underlying mode of action by which these variants cause or increase the disease risk is still obscure. Moreover, these variants collectively explain a minority (< 10%) of the total estimated genetic heritability of T2D (Voight et al., 2010). Some of this unexplained heritability has been attributed to the inefficiency of the current GWAS analysis to capture variants which exert small effects (Manolio et al., 2009). In order to increase the power to detect common genetic variants with modest effect, Voight et al. (2010) increased the sample size by performing a meta-analysis on eight T2D genome-wide association datasets, which discovered twelve novel risk loci for T2D. Amongst these newly identified loci, the genetic variants near the gene *KLF14* (Kruppel-like factor 14) displayed the most significant correlation with *KLF14* expression in adipose tissue (Voight et al., 2010), making it a strong candidate for further investigation. Not long after that, *KLF14* was also linked to high-density lipoprotein (HDL) cholesterol (Teslovich et al., 2010) and also believed to control the metabolic syndrome 'orchestra' (Civelek and Lusis, 2011).

KLF14, an imprinted gene (genes whose expression is dependent on the parent-of-origin) known to be expressed maternally, encodes for a transcription factor (TF) which acts as a master regulator of gene expression in adipose tissue (Small et al., 2011). Though KLF family genes are known to be involved in cell proliferation and differentiation (Dang et al., 2000), little is known about the function of *KLF14*. The GWASs have identified a set of non-coding SNPs in high linkage disequilibrium (LD), located upstream of *KLF14* (ranging between ∼4 kb to ∼48 kb upstream of its TSS) to be associated with T2D and HDL cholesterol (Small et al., 2011), collectively referred to as '*KLF14* locus' here (Fig. 2.1A). Interestingly, one of the SNPs in the *KLF14* locus (rs4731702, located 14 kb upstream of *KLF14*) was found to imitate the imprinting pattern of *KLF14* by correlating with the reduced expression of *KLF14* in the adipose tissue only when acquired maternally (Kong et al., 2009). This indicates *KLF14* to be the most likely gene influenced by the SNPs in the *KLF14* locus. These non-coding

SNPs in the *KLF14* locus potentially alter the *KLF14* expression by disrupting one or more *KLF14* associated regulatory elements such as enhancers, however SNPs in this locus which do so are not yet known. Furthermore, as *KLF14* is a TF regulating gene expression of other genes in adipose tissue, the SNPs affecting *KLF14* expression consequently have a *trans* effect on *KLF14* transcriptional targets, thus producing an assortment of metabolic defects (Small et al., 2011).

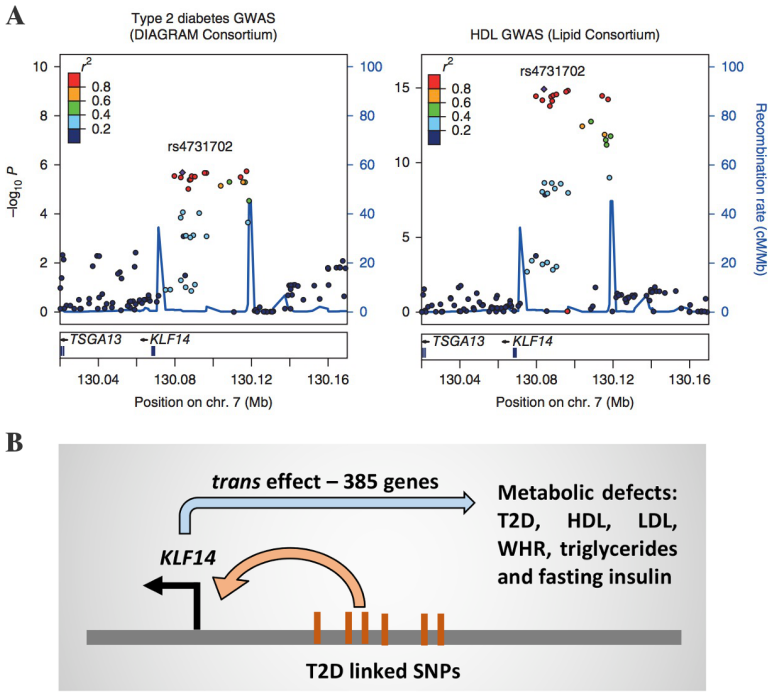


Fig. 2.1 Association of *KLF14* variants with metabolic traits. (A) The plots show the significance of the variants (as $-log_{10}p$) plotted against their genomic positions. Circles represent SNPs coloured according to their LD r^2 values with respect to the index SNP rs4731702. Figure taken from Small et al. (2011). (B) A schematic displaying the effect of T2D risk associated SNPs in the *KLF14* locus.

Although GWASs have identified that genetic variants in the *KLF14* locus are linked to T2D risk and HDL, it does not provide a biological explanation of how these SNPs exert the phenotypic effects, therefore, the mechanism by which the SNPs in the *KLF14* locus produce metabolic traits is not yet know. In order to gain insights into these mechanisms, the McCarthy and the Small lab investigated the functional network affected by the SNPs in the *KLF14* locus. They performed a meta-analysis by combining multiple genome-wide association traits which expanded the range of *KLF14* trait associations to low-density lipoprotein (LDL) cholesterol, triglycerides, waist-hip ratio and fasting insulin (Dupuis et al., 2010; Small et al., 2018). Furthermore, using expression data in adipose tissue from the TwinsUK cohort (Buil et al., 2015), they identified a *trans*-network of 385 genes affected by rs4731702 T2D risk allele in the *KLF14* locus (Fig. 2.1B). In order to identify if these affected genes are directly regulated by *KLF14*, they inspected the *KLF14* binding amongst them. Since the

binding sequence recognised and bound by *KLF14* has been previously identified using ChIP-seq (Jolma et al., 2013), the McCarthy lab scanned the upstream regions of genes in the *KLF14 trans*-network (up to 20 kb) for the enrichment of known *KLF14* binding site, and identified 177 (out of 385) genes to potentially harbour the *KLF14* binding site. This indicates that *KLF14* may directly regulate these 177 genes in the *trans*-network, while other genes may have a *KLF14* binding site further away from 20 kb or may be altered via an indirect effect (Small et al., 2018). Additionally, *SREBF1*, a *trans*-gene itself, was identified to regulate 18 other *trans*-genes in the network, 11 of which are not directly regulated by *KLF14*. *SREBF1* is an important TF in cholesterol homeostasis and appears to regulate a sub-network of genes involved in cholesterol biosynthesis and lipid metabolism (Small et al., 2018).

In a joint effort with the McCarthy lab, the Cox lab explored the metabolic effects of *Klf14* in the mouse genome. For this purpose, a mouse line featuring a complete loss of functional *Klf14* (*Klf14*$^{tm1(KOMP)Vlcg}$) was characterised by the Cox lab at the MRCHI. Mice were bred to form two cohorts: (1) heterozygotes with the deleted allele inherited maternally (MAT, equivalent to risk allele in human); and (2) heterozygotes with the deleted allele inherited paternally (PAT, equivalent to non-risk allele in human). Screening these mice through metabolic phenotyping tests revealed reduced HDL cholesterol in MAT males (Fig. 2.2A,E), along with a modest reduction in glucose homeostasis (Fig. 2.2B-D). Total and LDL cholesterol were also found to be relatively lower in MAT males (Fig. 2.2F-G), while no phenotypes were observed in the female mice (data not shown). However, no significant T2D traits were observed in these mice.

In collaboration with the Cox lab at the MRCHI, we hypothesised that the *KLF14* associated traits in humans may correlate better with the mouse at the molecular level, as opposed to the phenotypic level. Therefore, we sequenced the RNA from MAT and PAT mice with the aim to compare the *Klf14* associated transcriptional targets in mice with the *KLF14 trans*-network in humans. Here, using this RNA sequencing data, I describe the *Klf14* associated transcriptional network and phenotypes in the mouse. Since the T2D risk associated non-coding SNPs exert their effect via damaging regulatory elements, I performed a comparative epigenetic profiling of the human and mouse *KLF14* locus, with the goal to evaluate the relationship of active regulatory elements in this locus with T2D related pathways. I further analysed the human *KLF14* locus for phylogenetically conserved transcription factor binding sites (TFBSs) to identify potential *cis*-regulatory T2D risk variants which may affect *KLF14* expression. Finally, I investigated TF families for their preferential binding close to the T2D associated SNPs in the human locus. Overall, this study identifies potential enhancers, *cis*-regulatory variants and TFs associated with T2D risk in the *KLF14* locus, and provides insights

into the conserved and specific regulatory pathways associated with *KLF14* related phenotypes between the human and mouse genomes.

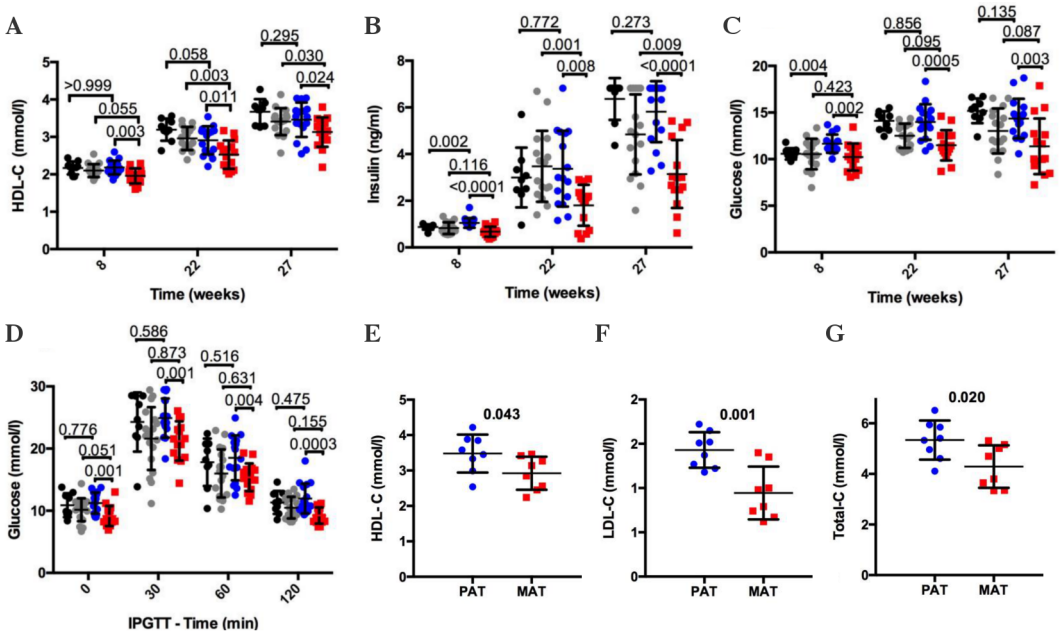


Fig. 2.2 Clinical chemistry analysis of the *Klf14* knockout mice. Plots display the clinical chemistry parameters measured between male *Klf14* knockout mice and their wild-type controls. All values are displayed as mean ± SD. For E-G, measurements were taken from blood sample of 33 week old mice. Grey: wild-type MAT; black: wild-type PAT; red: knockout MAT; blue: knockout PAT. Figure taken from Small et al. (2018).

2.2 Results

2.2.1 Transcriptional targets of *Klf14* in the mouse genome

To identify the genes regulated by *Klf14* in the mouse, we sequenced the RNA of subcutaneous fat taken from the $Klf14^{tm1(KOMP)Vlcg}$ MAT (n = 3) and PAT (n = 3) male mice. RNA-seq analysis comparing the global expression between MAT and PAT identified 285 differentially expressed genes at FDR < 0.05 (or 1599 genes at $p < 0.05$). Of these, 127 genes were up-regulated and 158 were down-regulated in the MAT mice, suggesting *Klf14* to act as both a repressor and an activator - a pattern also observed in the human *KLF14 trans*-network. The top differentially expressed genes with a FDR < 0.05 and fold change > ±2 (n = 60), along with their expression (in RPKM; reads per kilobase of transcript per million mapped reads) across all replicates are illustrated in Fig. 2.3. A Gene Ontology (GO) enrichment analysis shows that the genes influenced by the *Klf14* deletion (FDR < 0.05) are involved in biological processes relevant to the cholesterol phenotype observed in the knockout mice, such as response to lipid ($q < 7.6 \times 10^{-5}$),

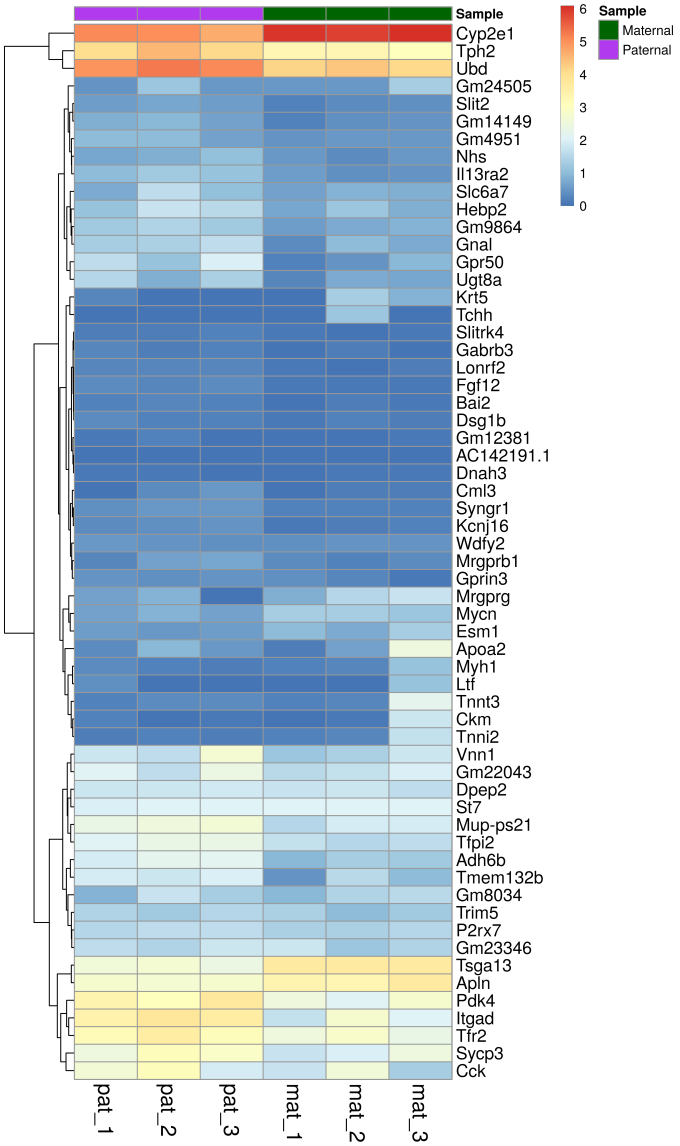


Fig. 2.3 Differentially expressed genes between *Klf14*$^{tm1(KOMP)Vlcg}$ PAT and MAT mice. Heatmap demonstrating the expression of significantly differentially expressed genes (FDR < 0.05, fold change > ±2, n = 60) across all biological replicates of PAT and MAT samples. The rows and columns of the heatmap represent genes and replicates respectively, while shading shows the level of gene expression in log RPKM. The rows of the heatmap are clustered using hierarchical clustering.

response to fatty acid ($q < 7.38 \times 10^{-3}$), regulation of metabolic processes ($q < 8.2 \times 10^{-4}$), response to steroid hormone ($q < 7.5 \times 10^{-3}$) and regulation of cell proliferation ($q < 1.4 \times 10^{-3}$) (Fig. 2.4). Furthermore, *Klf14* deletion also alters the expression of genes involved in the inflammatory response ($q < 2.2 \times 10^{-4}$), regulation of immune system process ($q < 9.7 \times 10^{-4}$) and leukocyte migration ($q < 1.2 \times 10^{-3}$), which agrees with earlier research proposing *Klf14* function in the regulation of T-regulatory cells (Sarmento et al., 2015).

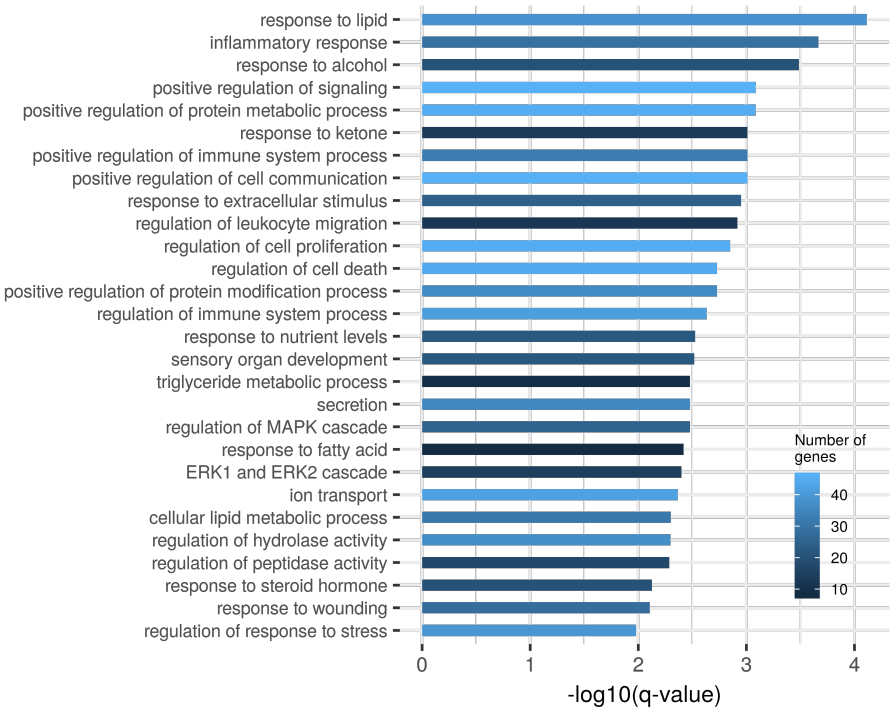


Fig. 2.4 GO enrichment analysis of *Klf14* transcriptional targets. Bar plot illustrating significantly over-represented GO terms amongst the differentially expressed genes between PAT and MAT mice (FDR < 0.05, n = 285). The enriched biological processes are displayed on the y-axis with their respective q-values (FDR) displayed on the x-axis. The gradient colour of the bars shows the number of differentially expressed genes related with the GO terms.

Next, I compared the *Klf14* transcriptional targets in the mouse to the human *trans*-network, which revealed a significant ($p < 10^{-6}$), but small number of genes common between them; 46 genes with differentially expressed gene set at $p < 0.05$, 8 genes with differentially expressed gene set at $q < 0.05$ respectively. This suggests that species specific *Klf14* targets are likely to be responsible for variable *Klf14* associated phenotypes between mouse and human. Though the majority of the *Klf14* transcriptional targets are different between mouse and human, I hypothesised that they may be engaged in the same functional pathways. A GO enrichment analysis of the genes in the human *trans*-network identified only 'oxidation-reduction process' ($q = 1.41 \times 10^{-2}$) as significant amongst the genes. Therefore, in order to check for possible interactions, I investigated the protein-protein interactions (PPIs) amongst the human *trans*-network and the mouse differentially expressed genes. This identified 255 targets (with 462 interactions) using the p-value gene set and 53 (with 105 interactions) using the q-value gene set, in the mouse to have a potential PPI with the mouse orthologous of the human *trans*-network genes (Fig. 2.5). Simulations conducted by adding random genes show that these interactions are significantly higher compared to what is expected by chance ($p \leq 0.0004$), suggesting that these *Klf14* targets in the mouse are likely to interact with their human *trans*-network partners compared to random protein-coding genes.

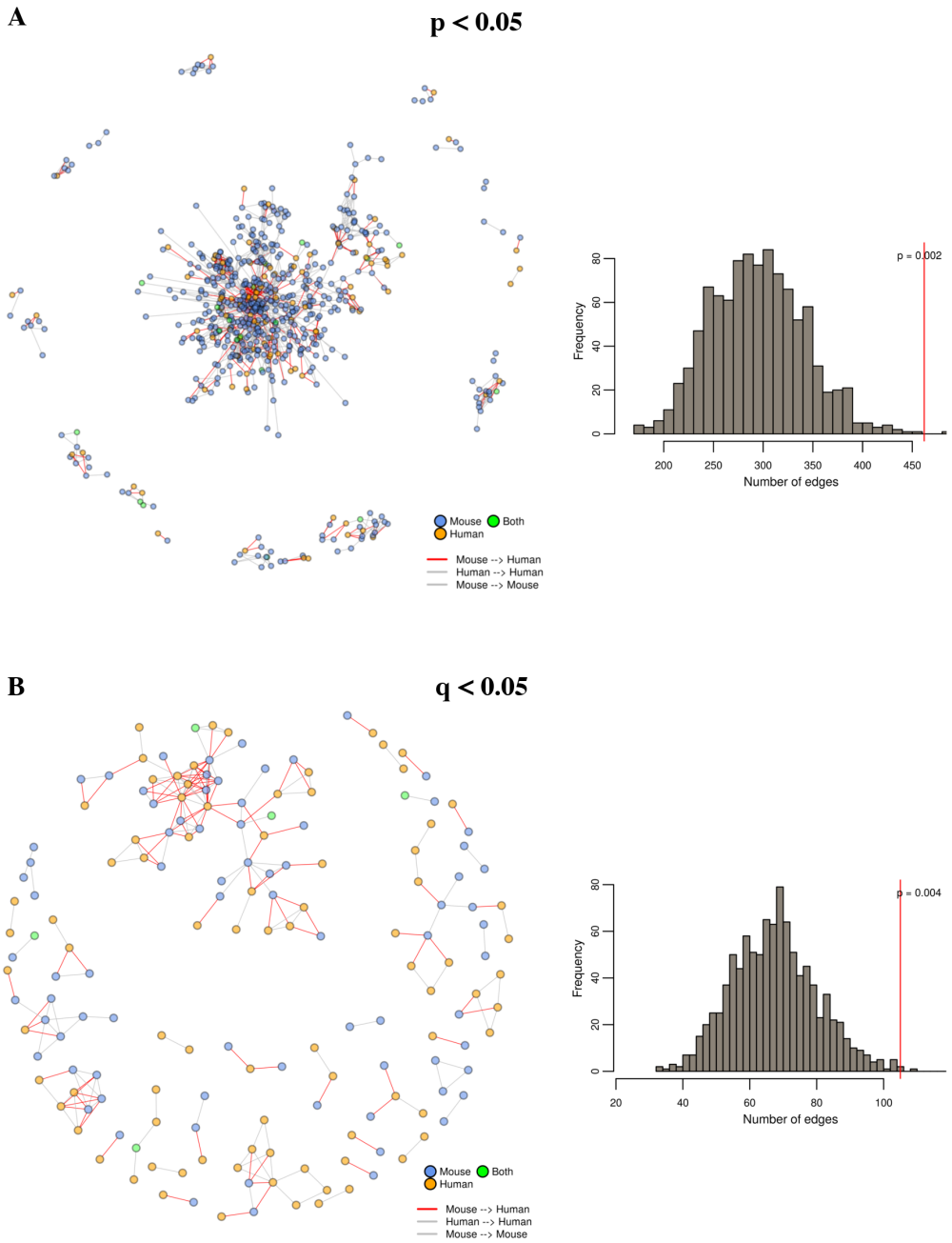


Fig. 2.5 Protein-protein interaction map amongst *Klf14* transcriptional targets in the mouse and the human *trans*-network genes. Networks (left panel) display potential PPIs between the human *trans*-network genes and the differentially expressed genes between MAT and PAT mice at (A) $p < 0.05$, and (B) $q < 0.05$. PPIs were retrieved from the STRING database (Franceschini et al., 2013) with the highest confidence (score > 0.9). Colour of the nodes signifies the source of the gene, and interactions (edges) between the mouse and human genes are highlighted in red. The histograms (right panel) show the results of the PPI network simulations. The red vertical line shows the observed number of edges between the mouse and human genes, and the grey distribution (obtained from 1,000 permutations) shows the number of edges between the mouse and randomly added protein-coding genes.

2.2.2 *De novo* motif discovery from *Klf14* transcriptional targets

In order to evaluate whether *Klf14* directly regulates the differentially expressed genes between mice expressing and not expressing *Klf14* (MAT and PAT respectively), I employed a *de novo* motif discovery approach to detect enriched motif sequences amongst the differentially expressed genes identified earlier. The strategy involved searching for over-represented motif sequences within upstream regions of differentially expressed genes (see methods 2.3.3). For input, I extracted two sets of regions upstream of differentially expressed genes: (1) a dataset including entire promoter sequence within an upstream distance of 300 bp, 500 bp, 750 bp and 1 kb, from the transcription start sites (TSSs) of differentially expressed genes; and (2) a dataset including genomic sequence from DNaseI hypersensitive sites (DHSs) in the fatpad tissue, within an upstream distance of 1 kb, 3 kb, 5 kb and 10 kb, from the TSSs of differentially expressed genes. To identify the over-represented motif sequences, I used two prominent community adopted tools for motif analysis, namely MEME (Bailey and Elkan, 1994a) and Homer (Heinz et al., 2010). Lastly, I performed this *de novo* motif analysis using two sets of genes: (1) differentially expressed genes identified at a statistical significance level of $p < 0.05$ (DEG_{pval}); and (2) differentially expressed genes identified at a FDR < 0.05 (DEG_{qval}).

For the majority of the sequence sets, the enriched motifs identified by both MEME and Homer were not highly significant, or included nucleotide repeats, and hence are likely to be false positives (Table 2.1 and 2.2, highlighted in grey). However, a highly significant motif was detected in DHSs 1 kb upstream of DEG_{pval} gene set (E-value $= 2 \times 10^{-43}$) (Table 2.2, highlighted in yellow). A similar motif, though not highly significant (E-value $= 2.3 \times 10^{-11}$), was also detected in promoter sequences 300 bp upstream of DEG_{pval} gene set (Table 2.1). Since the *Klf14* binding motif in the mouse genome is not known, I compared this *de novo* motif with the *KLF14* binding site in humans (Najafabadi et al., 2015). The *de novo* motif is strikingly similar to the human *KLF14* binding site ($q = 2.6 \times 10^{-3}$) and contains the CGCCC core, which is present in the binding sites of other Klf family members in the mouse such as *Klf1*, *Klf4*, *Klf5*, *Klf7*, *Klf9*, *Klf12*, *Klf13* and *Klf16* (Kulakovskiy et al., 2018; Mathelier et al., 2014) (Fig. 2.6). This shows that despite different transcriptional targets, a conserved regulatory motif between mice and human is involved in *Klf14* associated phenotypes.

The highly enriched motif sequence representing the potential binding site of *Klf14* in the mouse genome was identified in 142 transcriptional targets amongst DEG_{pval} (Fig. 2.7). This suggests that these genes may be directly regulated by *Klf14*, while the remaining genes may contain this *Klf14* motif outside the 10 kb upstream window scanned here or may be indirectly affected by the *Klf14* allele deletion via an interme-

Table 2.1 Over-represented motifs identified in promoter regions

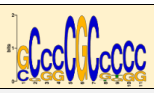
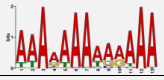
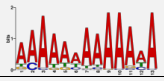
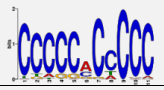
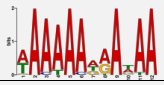
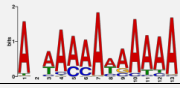
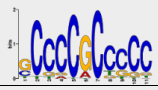
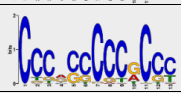
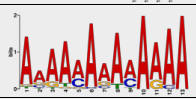
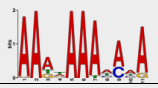
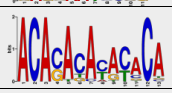
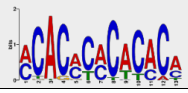
		Meme		Homer	
Gene set	Upstream Distance	E-value	Motif	P-value	Motif
DEG_{pval} (n= 1599)	300 bp	2.3e-11		1e-06	
	500 bp	1.8e-15		1e-07	
	750 bp	4.3e-28		1e-09	
	1 kb	1.5e-16		1e-09	
DEG_{qval} (n=285)	300 bp	1.6e-007		1e-07	
	500 bp	1.9e+00		1e-08	
	750 bp	7.2e-019		1e-10	
	1 kb	5.0e+004		1e-09	

Table 2.2 Over-represented motifs identified in upstream DHSs.

		Meme		Homer	
Gene set	Upstream Distance	E-value	Motif	P-value	Motif
DEG_{pval} (n= 1599)	1kb	2.0e-43		1e-08	
	3 kb	1.2e-07		1e-04	
	5 kb	7.4e-14		1e-05	
	10 kb	4.6e-05		1e-03	
DEG_{qval} (n=285)	1 kb	8.0e-05		1e-08	
	3 kb	7.2e+02		1e-07	
	5 kb	2.6e-03		1e-07	
	10 kb	9.9e-11		1e-06	

For Table 2.1 and Table 2.2: motifs highlighted in grey represent possible false-positives. E-value represents the probability of finding the same number of motifs with equal or higher log likelihood ratio in a set of equally sized random sequences.

diate pathway. Interestingly, similar to the human *trans*-network in which *SREBF1* is directly regulated by *KLF14*, *Srebf2* (mouse orthologue of *SREBF1*) is identified to be directly regulated by *Klf14* in the mouse network. *Srebf2*, a TF involved in the regulation of lipid and cholesterol homoeostasis, was differentially expressed between MAT and PAT mice ($p = 0.0006$). However, unlike the human *trans*-network where *SREBF1* was detected to regulate a sub-network of *trans* genes, *Srebf2* motif was not detected to be enriched in differentially expressed genes between MAT and PAT mice. I further scanned the promoter regions of these differentially expressed genes specifically for the presence of the human *SREBF1* motif and found no significant enrichment ($q \geq 0.163$). Nevertheless, *Srebf2* function appears to be one of the potential intermediate pathways affected by *Klf14* in both humans and mice.

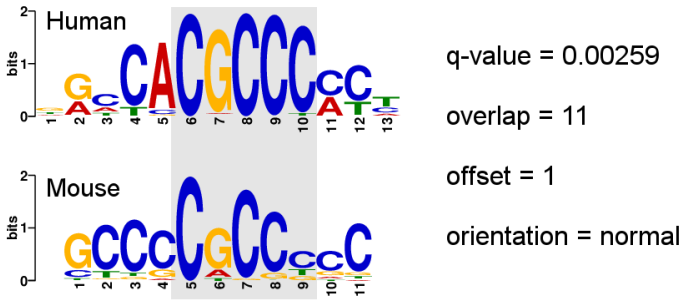


Fig. 2.6 Comparison of *Klf14* binding motif in human and mouse. Alignment of the human *KLF14* motif with the proposed binding site of *Klf14* in the mouse. The highlighted region in grey is the core of the binding site observed to be common in most of the Klf family members.

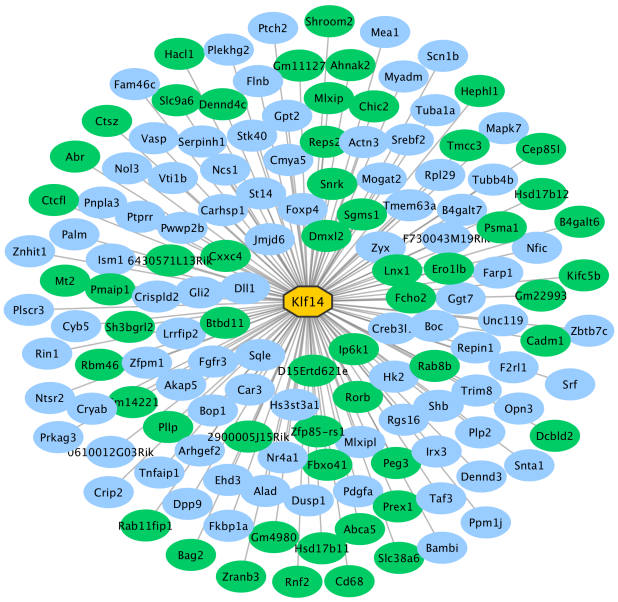


Fig. 2.7 Direct transcriptional targets of *Klf14* in the mouse. Network displaying the 142 differentially expressed genes enriched for the presence of a potential *Klf14* binding motif in the mouse. The nodes in the network are coloured to signify the direction of the effect in gene expression; ■ up-regulation, ■ down-regulation.

2.2.3 Epigenetic profiling of the *KLF14* locus

The *KLF14* locus in the human genome consists of a set of 23 non-coding T2D associated SNPs in high mutual LD. These SNPs potentially alter the *KLF14* expression via disrupting one or more regulatory elements associated with *KLF14*, therefore comparing the regulatory activity at *KLF14* locus between the human and mouse genomes may reveal insights about the common and species specific regulatory mechanisms related to *KLF14* function. To investigate regulatory activity at the human *KLF14* locus, the McCarthy lab used the chromatin state maps from the Roadmap Epigenomics Project (Roadmap Epigenomics Consortium et al., 2015). These chromatin state annotations were produced by integrating multiple histone mark data using a Hidden Markov Model (Ernst and Kellis, 2012). Comparing these annotations available in a wide range of tissues, the McCarthy lab identified a 1.6 kb long enhancer annotation ∼5 kb upstream of *KLF14* (chr7:130424000-130425600), which overlaps three of the T2D associated variants. It should be noted that the large size of this enhancer is maybe due to the way it was annotated. These chromatin state maps were produced using histone marks which often inflate the annotation lengths. This enhancer is adipose-specific, with a strong signal in adipose related cell-types such as adipose nuclei cells derived from adipose, mesenchymal stem cells derived from adipose, and adipocytes (Fig. 2.8A). There were no other adipose-specific enhancers detected in the *KLF14* locus.

Next, I inspected the region homologous to the human *KLF14* enhancer in the mouse genome. I asked the question whether this human *KLF14* enhancer region is active in the mouse? Also, whether any genomic changes in the mouse genome, possibly at positions homologous to the human SNPs would affect the expression of *Klf14* in the mouse? To investigate these questions, I scanned the *Klf14* locus in the mouse for any evidence of an active enhancer similar to what is observed in the human genome. Since no chromatin state or enhancer annotations were available for the mouse genome at the time of analysing, I used DHSs from ENCODE to explore open chromatin sites which might indicate enhancer activity. The human *KLF14* enhancer mapped to a homologous region in the mouse ∼4.5 kb upstream of *Klf14* (Fig. 2.8B). However, no DHSs were detected in any of the tissues within this region, suggesting that this region in the mouse does not exhibit regulatory activity. Furthermore, the homologous T2D associated SNP positions did not overlap any significant DHSs in the mouse tissues. This indicates that targeting the homologous T2D associated SNP positions in the mouse might not alter the *Klf14* expression in the mouse genome. Hence, mouse models using CRISPR-Cas9 to create mutations at these positions are likely not to produce the same phenotype effect as observed in the humans.

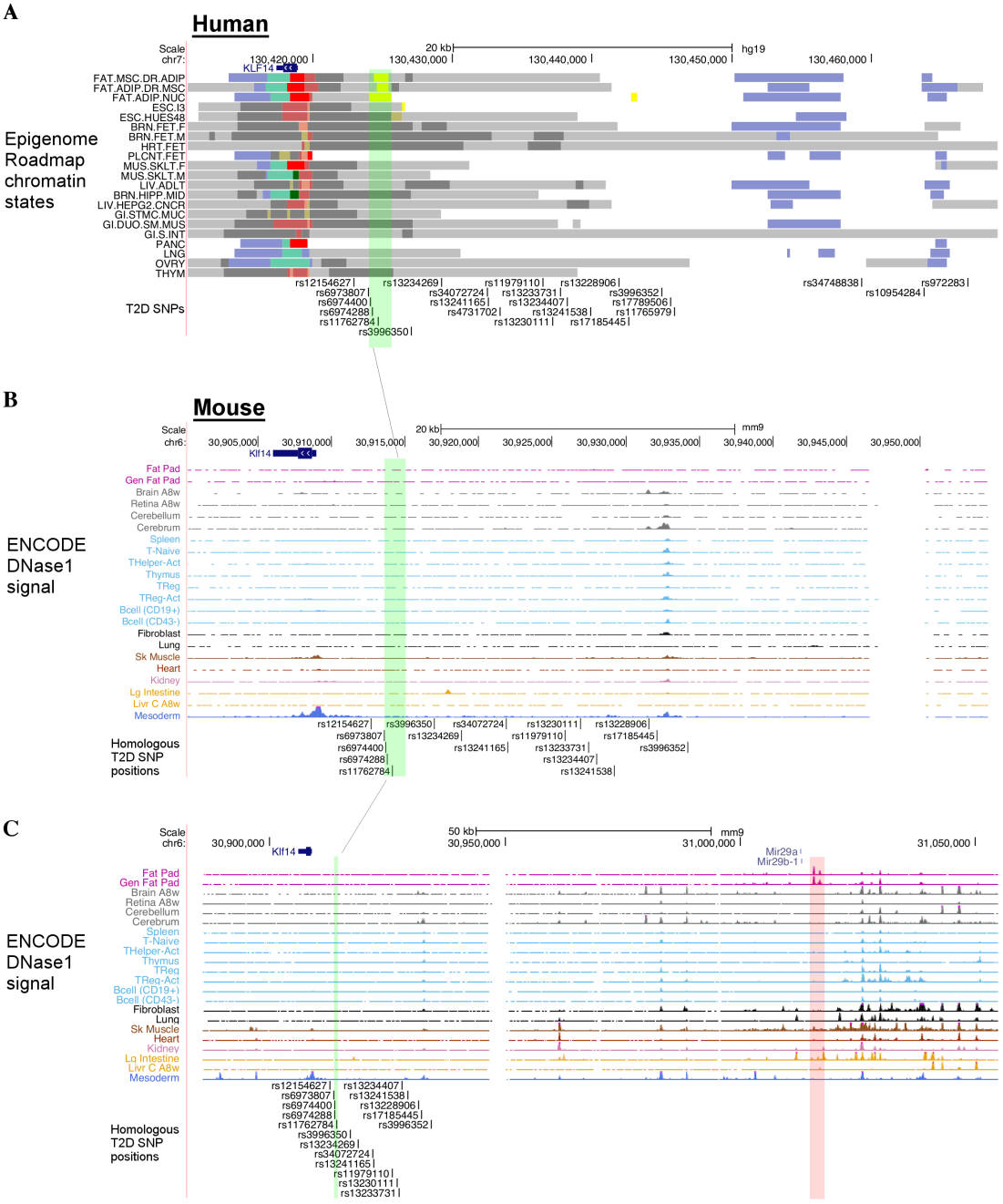


Fig. 2.8 Comparison of *KLF14* epigenomic landscape between human and mouse genomes. Genome browser snapshot of *KLF14* locus in the (A) human and (B-C) mouse genomes. The human locus displays the chromatin state annotations in different tissues from the Roadmap Epigenomics Project and SNPs associated with T2D risk. The mouse locus displays the DHSs in different tissues from ENCODE and homologous positions of T2D associated SNPs. The region highlighted in green shows the *KLF14* associated enhancer in human (A) and its homologous location in the mouse genome (B, C), and the region highlighted in red shows a highly enriched fat pad specific DHS identified in the mouse (C). Chromatin state annotations: ■ Active TSS; ■ Enhancers; ■ Weak repressed polycomb; ■ Repressed polycomb; ■ Heterochromatin; ■ ZNF genes & repeats; ■ Bivalent/poised TSS.

There were no other DHSs observed in the fat pad within the proximal *Klf14* locus, though high DNaseI enrichment was detected ∼106 kb upstream of *Klf14*, with high specificity to the mouse fat pad and the genital fat pad (Fig. 2.8C). Apparently, this ∼300 bp long DHS is near the Mir-29 family of microRNAs (*Mir29a* and *Mir29b-1*), which are highly expressed in pancreas and liver, and have been recognised to regulate insulin signalling and glucose homeostasis in mouse models of obesity and diabetes (Dooley et al., 2016; He et al., 2007; Massart et al., 2017; Yang et al., 2014). Thus, this DHS may be associated with the Mir-29 microRNAs and cannot be directly associated with *Klf14* without any other functional data.

Finally, I analysed the human *KLF14* enhancer to identify what potential TFs bind there and if their function is related to T2D pathways (Table. 2.3). Since no active enhancers were identified in the mouse *Klf14* locus, this analysis was not performed in the mouse genome. I scanned the human *KLF14* enhancer region using FIMO for all the known motifs currently available in the open source motif databases (see methods 2.3.4). Potential TFs binding within the human *KLF14* enhancer involves regulators known to be involved in adipose biology and diabetes, such as *HMGA2*, known to play a role in preadipocyte proliferation (Xi et al., 2016) and associated with T2D (Markowski et al., 2013; Voight et al., 2010); *TFL2* (also known as *TCF7L2*), previously connected with the T2D risk (Florez, 2007; Grant et al., 2006; Groop, 2010; Groves et al., 2006; Lyssenko et al., 2007); IRF family members (*IRF1, IRF2, IRF3, IRF4, IRF5*), which acts as regulators of adipogenesis (Eguchi et al., 2008; Kumari et al., 2016); STAT family members (*STAT1, STAT2, STAT5A, STAT5B, STAT6*), involved in maintenance of adipocytes (Harp et al., 2001; Stephens et al., 1996); and HNF1, responsible for Maturity-onset diabetes of the young (MODY) - a monogenic form of diabetes mellitus (Ellard and Colclough, 2006; Frayling, Bulamn, et al., 1997; Frayling, Bulman, et al., 1997). However, these potential TFBSs did not overlap any of the T2D associated SNPs within the *KLF14* enhancer (Fig. 2.9). This could be due to the limited number of motif PWMs in the public domain used in this analysis and also because our current knowledge about TFs is incomplete. Overall, this data shows that TFs involved in adipose biology and diabetes which bind within the human *KLF14* associated enhancer are potentially missing regulatory activity in the corresponding homologous mouse region, suggesting that *KLF14* associated enhancer regions and their underlying regulatory mechanisms in the *KLF14* locus might have diverged between the human and mouse genomes.

Table 2.3 Potential TF PWM matches (q < 0.01) in the human *KLF14* associated enhancer (chr7:130424000-130425600).

Motif	Start	End	Strand	p-value	q-value	Matched sequence
TF7L2	148	159	+	7.81E-07	0.000378	AACATCAAAGAG
IRF3	35	51	+	1.23E-06	0.00058	TGAAAAGGAAACTAGAA
HMGA2	19	33	+	2.30E-06	0.000733	ATAATTGGGGATTAT
STAT2	35	49	+	1.99E-06	0.000943	TGAAAAGGAAACTAG
HMGA2	20	34	-	6.11E-06	0.000973	AATAATCCCCAATTA
NR5A2	83	92	-	1.78E-06	0.00104	TTCAAGGCCA
PITX2	312	321	-	2.17E-06	0.00107	TGGGATTAAT
ISRE	34	48	-	3.46E-06	0.00166	TAGTTTCCTTTTCAA
ARI3A	287	308	-	8.26E-06	0.0018	CATAAATGAATAGTAATACTAA
STA5A	335	346	+	4.60E-06	0.00203	CATTCCTAGAAA
STAT5B	334	348	-	8.41E-06	0.00283	TATTTCTAGGAATGT
STAT5B	334	348	+	1.15E-05	0.00283	ACATTCCTAGAAATA
STA5B	335	347	-	6.38E-06	0.00308	ATTTCTAGGAATG
HAND1	7	18	+	5.36E-06	0.0031	GGGTCTGGAAGT
HNF1	304	318	-	1.14E-05	0.00319	GATTAATAATCATAA
HMGA2	19	33	-	3.23E-05	0.00343	ATAATCCCCAATTAT
STA5B	335	347	+	1.72E-05	0.00416	CATTCCTAGAAAT
BCL6	330	345	+	1.04E-05	0.00423	TTAAACATTCCTAGAA
BCL6	337	352	-	1.87E-05	0.00423	GCAGTATTTCTAGGAA
IRF4	36	51	+	9.05E-06	0.00436	GAAAAGGAAACTAGAA
ARI3A	288	309	-	4.24E-05	0.00462	TCATAAATGAATAGTAATACTA
HNF1	56	70	+	3.35E-05	0.00468	GGTAATTTTTTAATG
HMGA2	50	64	+	5.96E-05	0.00475	AATAAGGGTAATTTT
HNF1A	304	318	-	1.58E-05	0.00498	GATTAATAATCATAA
IRF2	35	48	+	1.12E-05	0.00535	TGAAAAGGAAACTA
GATA4	110	118	-	1.20E-05	0.00553	AGAGATAAC
STAT5A	334	348	+	1.09E-05	0.00562	ACATTCCTAGAAATA
STA5A	336	347	-	2.58E-05	0.00571	ATTTCTAGGAAT
STAT1	334	348	-	1.14E-05	0.00593	TATTTCTAGGAATGT
STF1	83	91	-	1.15E-05	0.0068	TCAAGGCCA
ARI3A	17	38	+	9.37E-05	0.00681	GTATAATTGGGGATTATTTGAA
IRF1	36	47	+	1.55E-05	0.00738	GAAAAGGAAACT
STAT6	337	349	+	1.91E-05	0.00756	TTCCTAGAAATAC
POU3F2	60	73	-	2.89E-05	0.00781	CTGCATTAAAAAAT
HNF1A	56	70	+	5.07E-05	0.00796	GGTAATTTTTTAATG
IRF8	34	48	+	1.53E-05	0.00802	TTGAAAAGGAAACTA
EHF	379	391	-	1.57E-05	0.00816	AAGGCAGGAAGGA
IRF5	94	113	+	2.02E-05	0.00833	TAAATGTTACCAAAATGTTA
IKZF1	116	123	-	1.51E-05	0.00898	TTGGGAGA
MEF2	54	69	-	2.33E-05	0.00904	ATTAAAAAATTACCCT
MEF2	26	41	-	4.34E-05	0.00904	CTTTTCAAATAATCCC
OCT1	289	311	-	4.86E-05	0.00906	AATCATAAATGAATAGTAATACT
OCT1	298	320	+	5.70E-05	0.00906	ATTCATTTATGATTATTAATCCC
HXC6	299	313	+	2.83E-05	0.00957	TTCATTTATGATTAT

Table lists the significant PWM matches (sorted by q-value) in the human *KLF14* associated enhancer analysed using FIMO (Grant et al., 2011). The start and end positions of the motif matches are relative to the enhancer start region. The p-value represents the probability of a random DNA sequence of the same length as the motif with as good as or better sequence match score than the motif at that position. The q-value of a motif occurrence represents the false discovery rate if the occurrence is accepted as significant.

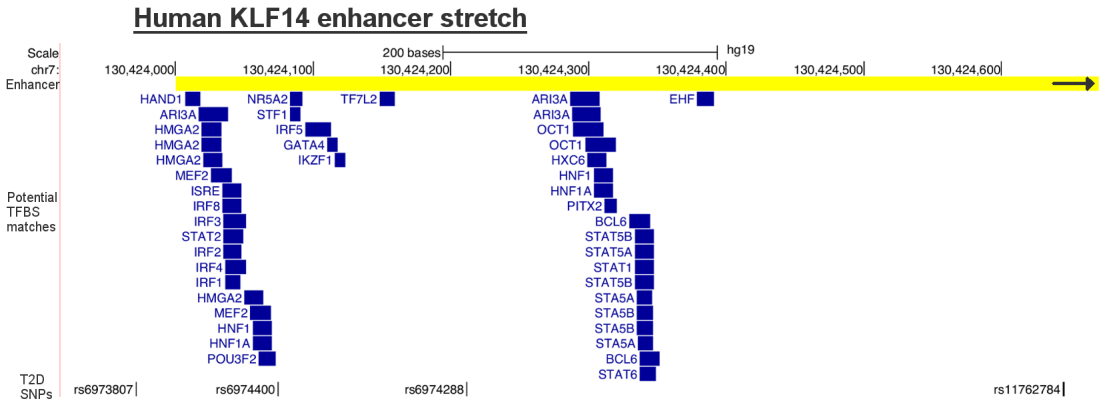


Fig. 2.9 Genomic view of the potential motif matches in the human *KLF14* associated enhancer. Genome browser snapshot of the human *KLF14* associated enhancer displaying possible TF PWM matches (FIMO $q < 0.01$) and T2D risk associated SNPs associated.

2.2.4 Phylogenetic Module Complexity Analysis

The regulatory regions in eukaryotes have a tendency to exist in *cis*-regulatory modules (CRMs), which consist of combinations of TFs binding together. These CRMs usually situated upstream of the TSSs, regulate expression of genes and any changes in the TFBSs within the CRMs could lead to phenotypic changes. Therefore, non-coding variants occurring within or near the CRMs could potentially modulate gene expression and contribute to disease genetics. In order to identify potential *cis*-regulatory variants in the human *KLF14* locus, I implemented a method called Phylogenetic Module Complexity Analysis (PMCA), outlined in Claussnitzer et al. (2014). PMCA investigates each non-coding variant by scanning its ± 60 bp flanking region for TF binding patterns conserved across species, to predict *cis*-regulatory variants occurring in regions of notably high TF binding complexes. For each flanking sequence, orthologous sequences are identified across multiple vertebrate species, which are further searched to detect conserved TFBSs, TFBS modules and TFBSs within those modules (see methods 2.3.5 for the pseudo code). A module is defined as 'complex' comprising of two or more conserved TFBSs appearing in the same order within a distance range, in all or subset of the orthologous sequences. This method eventually classifies the region around the non-coding variant to be complex, i.e. significantly enriched in conserved TFBS patterns, or 'non-complex' if the occurrences of TFBSs is less than what is expected by random. The expected enrichment of conserved TFBSs is estimated by randomising the sequences in the original set of orthologous sequences.

I applied PMCA to the set of 23 T2D associated non-coding variants in the *KLF14* locus. PMCA predictions suggest 16 variants to be in complex regions and 7 in non-complex regions (Fig. 2.10 and Table. 2.4). The SNPs rs6974400, rs4731702 and rs12154627 were identified to be the top three complex SNP regions based on the

number of TFBSs in conserved TFBS modules (Table. 2.4). Next, I investigated the enrichment of TFs binding around these complex SNP regions, as the TFs recruited here could give insights into the mechanistic action of these *cis*-regulatory variants. To detect potential TF binding near the SNPs, I implemented a previously described positional bias algorithm (Hughes et al., 2000), which measures the tendency of a TF to preferentially bind within a particular distance to the *cis*-regulatory variants. The $\pm$500 bp sequences flanking the SNPs in complex and non-complex regions were scanned to detect potential binding of TF families (314 TF PWMs grouped into 100 TF families). The sequences were searched within a sliding window of 50 bp (with incremental steps of 10 bp) and positional bias score for each TF family was calculated as the binomial probability to obtain observed number of matches within the window out of a possible total matches in the sequence (see methods 2.3.6 for the pseudo code).

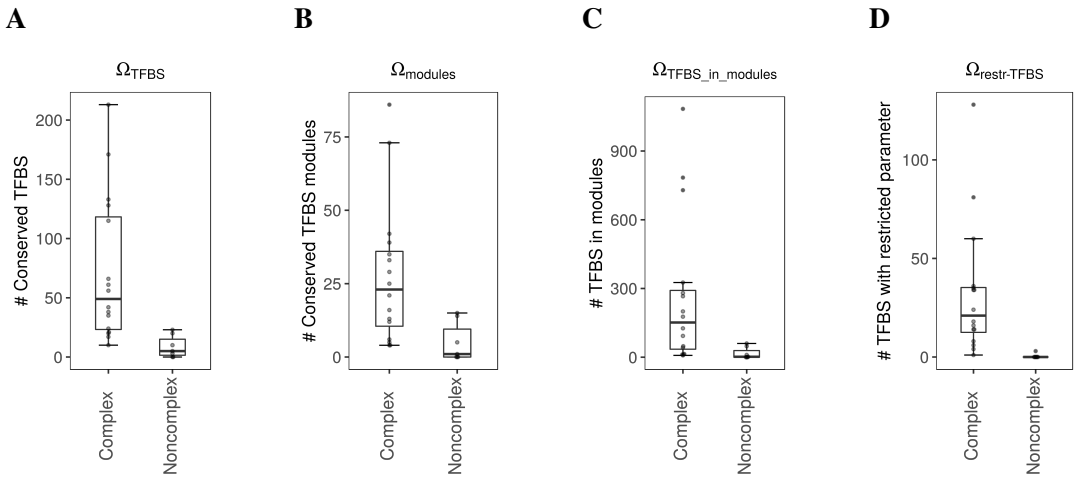


Fig. 2.10 Identification of *cis*-regulatory variants in the *KLF14* locus. Box plots showing the occurrences of (A) conserved TFBSs Ω_{TFBS}, (B) conserved TFBS modules $\Omega_{modules}$, (C) conserved TFBSs in modules $\Omega_{TFBS\_in\_modules}$, and (D) TFBSs counted in at least half of the orthologous sequences $\Omega_{restr-TFBS}$ for complex and non-complex SNP regions. The number of TFBSs were calculated by repeated counting with different number of sequences in the orthologous sequence set to weigh the extent of conservation in TFBSs (see methods 2.3.5 for details). Each box plot shows: the median, middle bar; interquartile range, the box; 1.5 times the interquartile range, the whiskers.

Table 2.4 Classification of T2D associated variants in the *KLF14* locus using the PMCA

SNP	Ω_{TFBS}	$\Omega_{modules}$	$\Omega_{TFBS\_in\_modules}$	$\Omega_{restr-TFBS}$	S_{all}	PMCA result
rs10954284	20	4	8	6	9	Complex SNP region
rs11762784	10	5	10	0	9	Non-complex SNP region
rs11765979	56	35	280	34	9	Complex SNP region
rs11979110	213	25	177	128	7.37	Complex SNP region
rs12154627	133	73	729	14	9	Complex SNP region
rs13228906	3	1	2	0	8.4	Non-complex SNP region
rs13230111	115	33	200	60	9	Complex SNP region
rs13233731	24	21	93	14	9	Complex SNP region
rs13234269	0	0	0	0	0	Non-complex SNP region
rs13234407	10	5	14	4	9	Complex SNP region
rs13241165	35	12	42	18	9	Complex SNP region
rs13241538	17	4	12	8	9	Complex SNP region
rs17185445	5	0	0	3	3	Non-complex SNP region
rs17789506	38	29	266	16	9	Complex SNP region
rs34072724	42	13	48	24	9	Complex SNP region
rs34748838	21	16	126	1	7.96	Complex SNP region
rs3996350	20	14	48	0	9	Non-complex SNP region
rs3996352	61	39	326	35	9	Complex SNP region
rs4731702	171	42	784	81	7.72	Complex SNP region
rs6973807	0	0	0	0	0	Non-complex SNP region
rs6974288	23	15	60	0	9	Non-complex SNP region
rs6974400	128	86	1084	34	9	Complex SNP region
rs972283	66	6	12	36	7.7	Complex SNP region

For the complex SNP regions, a significant positional bias around the SNPs ($\pm$25 bp) was detected for the forkhead family (FKHD) of TFs (Fig. 2.11). This preferential binding was not detected to be as strong at the SNPs in non-complex regions. However, an inspection of the SNPs loci in complex regions revealed that none of the TFs in FKHD family overlaps the exact position of the SNPs. Despite this, forkhead TFs namely *FOXP3*, *FOXJ3* and *FOXO1* were detected to potentially bind within $\pm$25 bp of the variants rs11979110, rs17789506 and rs4731702 respectively. The FOX genes have been identified to have numerous important biological functions including a role in insulin signalling (Hannenhalli and Kaestner, 2009; Yang et al., 2009). Especially, *FOXO1*, which is involved in insulin secretion and glucose production (Dong et al., 2008; Nakae et al., 2002; Nakae et al., 2008). Consequently, *FOXO1* could conceivably be involved in the underlying mechanisms by which T2D risk variants in the *KLF14* locus produce metabolic traits.

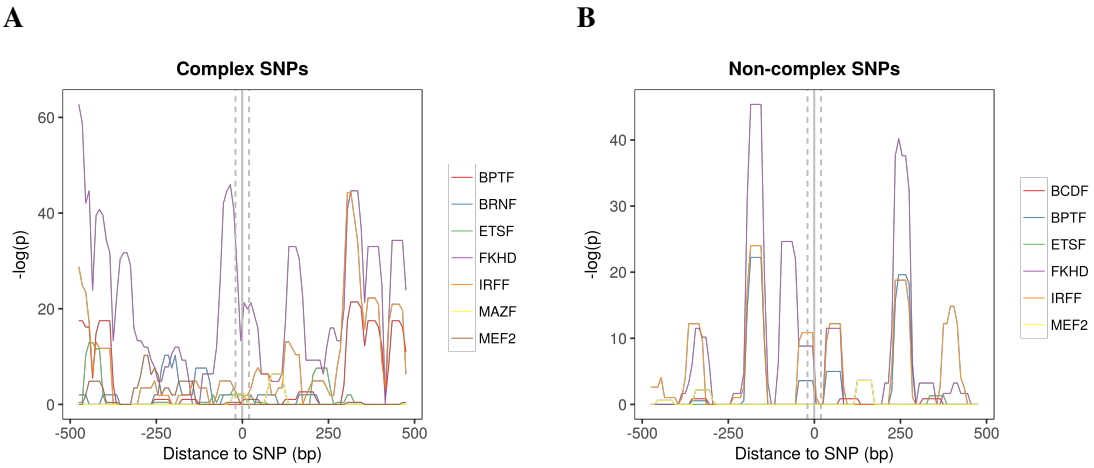


Fig. 2.11 TF binding positional bias with respect to the T2D SNPS in the *KLF14* locus. Distribution of TFBS occurrences of various TF families relative to the T2D SNP positions in the *KLF14* locus. The region around (A) complex and (B) non-complex SNPs (±500 bp) was searched in windows of 50 bp for potential TFBS matches and their preferential binding positions using the positional bias analysis. A strong enrichment of FKHD TFs is observed within ±25 bp (grey dashed lines) of the T2D SNPs in the complex regions.

2.3 Methods

2.3.1 Datasets

The set of variants associated with T2D in the *KLF14* locus were retrieved from Voight et al. (2010) and their hg19/GRCh37 genome coordinates were used for all the analysis. For investigating the epigenetic activity within the human *KLF14* locus, primary 15 state chromHMM annotations from the Roadmap Epigenomics Project were used by the McCarthy lab. These annotations were accessed and visualised using the Roadmap Epigenomics data public hub on the UCSC genome browser. The Roadmap Epigenomics Project dataset consists of three tissues related to adipose, namely adipose nuclei cells derived from adipose, mesenchymal stem cells derived from adipose, and adipocytes. Since no chromHMM annotations were available for the mouse genome, DHS data from the ENCODE (University of Washington) was used as an indicator for potential distal enhancer activity. This ENCODE dataset consists of mouse adipose tissue derived from the fat pad and the genital fat pad.

2.3.2 RNA-seq data analysis

The RNA-seq reads were aligned to the mouse genome (mm10) using TopHat (Trapnell et al., 2012). The read counts for each gene were calculated from the aligned reads using HTSeq (Anders et al., 2015), and differentially expressed genes were identified using three tools; EdgeR (Robinson et al., 2010), DESeq (Anders and Huber, 2010)

and Cufflinks (Trapnell et al., 2012), default parameters were used for all of them. The raw counts in each sample were used to calculate the RPKM (Reads per million per kilo-base) and were visualised in a heatmap using R. The GO enrichment analysis for differentially expressed genes was performed using ToppFun (Chen et al., 2009) and the enriched GO terms with a FDR < 0.05 were considered significant.

2.3.3 *De novo* motif discovery

To identify over-represented motif sequences, the entire promoter and DHSs upstream of differentially expressed genes were scanned. The DHSs were retrieved from the fat pad tissue available in the mouse ENCODE dataset. Multiple upstream distance windows were used and every case was run individually and independently. The upstream sequences were extracted from the mouse genome (mm9) using BEDTools (Quinlan, 2014). Two different tools, namely MEME (Bailey and Elkan, 1994a) and Homer (Heinz et al., 2010), were employed to detect *de novo* motifs enriched amongst the differentially expressed genes. A workflow of the de novo motif strategy used here is described in Fig. 2.12. This analysis was performed on two sets of genes: (1) differentially expressed genes identified at a significance level of $p < 0.05$ (DEG_{pval}), and (2) differentially expressed genes identified at FDR < 0.05 (DEG_{qval}). Due to the high rate of false-positives in motif enrichment analysis, only highly significant motifs ($p \leq 10^{-30}$) were considered for further analysis.

2.3.4 Identifying known transcription factor binding sites

In order to scan a genomic sequence for potential TFBSs of known motifs, the tool FIMO (Find Individual Motif occurrences) (Grant et al., 2011) from the MEME suite was employed. FIMO requires a database of motif sequence PWMs for known TFs from the user. Each TF motif from the database is treated independently and searches for its occurrences in the provided sequence. A custom database of motif PWMs from publicly available JASPAR (Mathelier et al., 2014) and TRANSFAC (Wingender et al., 1996) motifs was constructed for this purpose, which comprised of 554 and 398 motifs for the human and mouse genome respectively. FIMO estimates a p-value for each motif which describes the probability of a random sequence of the same length to achieve same or higher score at that particular position, and further calculates the false discovery rate (q-value). To reduce false positives, only the motif matches achieving a q-value < 0.01 were considered for further analysis.

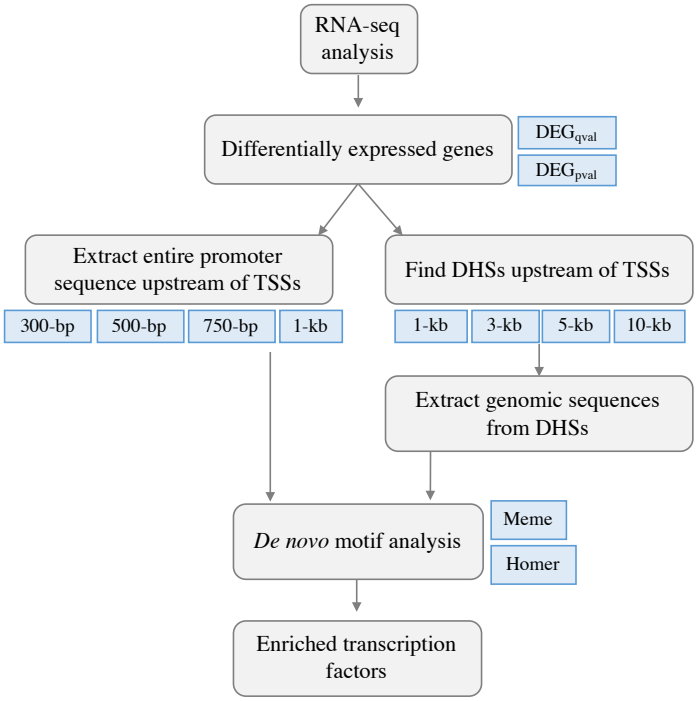


Fig. 2.12 *De novo* motif discovery strategy. Workflow demonstrating the *de novo* motif discovery approach employed to search enriched motif sequences upstream of differentially expressed genes.

2.3.5 Phylogenetic Module Complexity Analysis

A previously described phylogenetic module complexity analysis (PMCA) algorithm was used to identify potential *cis*-regulatory variants within the CRMs. The PMCA method described in Claussnitzer et al. (2014) uses commercially available Genomatix software suite for the analysis and commercial TF databases like TRANSFAC for the TF PWMs. Therefore, a modified version of PMCA using open source tools and TF databases was written in perl by me. For each SNP, 60 bp surrounding sequence on each side was retrieved (hg38/GRCh38) and orthologous sequences to this region (if any) in 13 mammalian species were extracted from Ensembl multiple alignments (EPO set, version 82). FIMO was then used to identify potential TFBS matches in these sequences. TFBS matches conserved across all or subset of the orthologous sequences were identified and TFBSs conserved in the same order within a distance range were defined into conserved TFBS modules. These phylogenetically conserved TFBSs and modules were repeatedly counted to estimate the degree of conservation, and the occurrences of TFBSs (Ω_{TFBS}), TFBS modules ($\Omega_{modules}$) and TFBSs in modules ($\Omega_{TFBS\_in\_modules}$) were calculated. To estimate the background probabilities, these steps were repeated 1,000 times using randomly shuffled sequences from the orthologous set, and the number of conserved TFBSs ($\Omega_{TFBS^{rnd}}$), TFBS modules ($\Omega_{modules^{rnd}}$) and TFBSs in modules

($\Omega_{TFBS\_in\_modules^{rnd}}$) occurring by random chance were calculated. The sequences were locally shuffled in windows of 10 bp to preserve the local nucleotide frequency. Finally, based on the observed and expected number of conserved TFBS patterns, an overall score (S_{all}) was calculated which classified the input non-coding variant into a complex or a non-complex SNP region. The pseudo code of the modified PMCA is outlined as below:

Algorithm 1 PMCA - Phylogenetic Module Complexity Analysis

for each non-coding SNP **do**
  Select the region around the SNP
    start position = SNP position - 60 bp
    end position = SNP position + 60 bp
  Extract orthologous sequences
    Extract orthologous sequences from multiple alignments of 13 mammals using Ensembl API (set S)
  Calculate modular complexity
    for each sequence set S **do**
      $N_S \leftarrow$ number of sequences in S
      Run FIMO to identify potential known TFBSs
      Calculate conserved TFBSs in S ($q < 0.01$)
      Define TFBSs in modules
      for $i = 2$ to N_S **do**
        ω_{TFBS} = number of conserved TFBSs in atleast i sequences of S
        $\Omega_{TFBS} = \Omega_{TFBS} + \omega_{TFBS}$
        $\omega_{restr-TFBS}$ = number of conserved TFBSs in atleast $\frac{i \times 100}{N_S}$ sequences of S
        $\Omega_{restr-TFBS} = \Omega_{restr-TFBS} + \omega_{restr-TFBS}$
        $\gamma \leftarrow$number of conserved TFBSs required in a module to be counted
        for $\gamma = 2$ to 10 **do**
          $\omega_{\gamma-modules}$ = number of modules with γ TFBSs in atleast i sequences of S
          $\omega_{TFBS\_in\_\gamma-modules}$ = number of TFBSs in modules with γ TFBSs in atleast i sequences of S
          $\Omega_{modules} = \Omega_{modules} + \omega_{\gamma-modules}$
          $\Omega_{TFBS\_in\_modules} = \Omega_{TFBS\_in\_modules} + \omega_{TFBS\_in\_\gamma-modules}$
        end for
      end for
    end for
  Randomly shuffle sequence set S- repeat 1,000 times
    permute the bases in each 10 bp window to generate a randomised sequence set similar in local nucleotide distribution to S
    calculate modular complexity to obtain $\Omega_{TFBS^{rnd}}$, $\Omega_{restr-TFBS^{rnd}}$, $\Omega_{modules^{rnd}}$, and $\Omega_{TFBS\_in\_modules^{rnd}}$
  Estimate probabilities
    $p-est_{TFBS} = f(\Omega_{TFBS^{rnd}} \geq \Omega_{TFBS})$
    $p-est_{restr-TFBS} = f(\Omega_{restr-TFBS^{rnd}} \geq \Omega_{restr-TFBS})$
    $p-est_{modules} = f(\Omega_{modules^{rnd}} \geq \Omega_{modules})$
    $p-est_{TFBS\_in\_modules} = f(\Omega_{TFBS\_in\_modules^{rnd}} \geq \Omega_{TFBS\_in\_modules})$

if $count(\Omega_{irnd} \geq \Omega_i) = 0$ **then**
$p - est_i = 1/1001$
where $i \leftarrow$ TFBS, rest-TFBS, modules or TFBS_in_modules
end if
Calculate overall score
$S_{all} = -log(p - est_{TFBS} \times p - est_{modules} \times p - est_{TFBS\_in\_modules})$
Classify the non-coding SNP
if $S_{all} > 6.5$ & $p - est_{restr-TFBS} < 0.15$ & $p - est_{TFBS} < 0.075$ **then**
SNP $\leftarrow$ complex region
else
SNP $\leftarrow$ non-complex region
end if
end for

2.3.6 Positional bias algorithm

The positional bias analysis (Hughes et al., 2000) was employed to further identify the TFs with preferential binding patterns at T2D SNP positions. The $\pm$500 bp sequences flanking the set of T2D SNPs were scanned to identify potential binding of TF family PWMs using FIMO. For this analysis, 314 TF PWMs were grouped into 100 TF families, as searching for a TF family instead of the individual TFs removes redundancy and identifies the best match within a TF family. The TF family motifs were searched using a sliding window of 50 bp, with a 10 bp step increase after every iteration. The positional bias binomial probability for each TF family and each window was estimated using the formula:

$$P = \sum_{i=m}^{t} (\frac{t}{i})(\frac{w}{l})^i (1 - \frac{w}{l})^{t-i} \tag{2.1}$$

where m is the exact number of TF family matches within the scan window, t is the total number of matches in the sequence, w is the size of the scan window (i.e. 50) and l is the length of the sequence (i.e. 1,000). For visualisation of the results, -$\log_{10}$(p) and the distance of the SNP from the middle of the scan window was calculated. This positional bias algorithm using open source tools and public TF databases was written in perl and its pseudo code is described below:

Algorithm 2 Positional Bias Analysis

for each SNP in complex/non-complex set **do**
Select the region around the SNP
start position = SNP position - 500 bp
end position = SNP position + 500 bp
extract FASTA genomic sequence
Identify TFBSs

use FIMO to identify potential TF family PWM matches ($q < 0.01$)
end for
Calculate positional bias - binomial probability
for each TFBS family **do**
calculate $t \leftarrow$ total number of matches in the sequences
set window start to 0
set window size (w) to 50 bp
set sequence length (l) to 1,000 bp
for window start = 0 to 950 **do**
calculate $m \leftarrow$ number of TFBS matches within this window
Positional Bias, $P = \sum_{i=m}^{t} (\frac{t}{i})(\frac{w}{l})^i(1-\frac{w}{l})^{t-i}$
increment window start by 10 bp
end for
end for
END

2.4 Discussion

T2D associated variants in the *KLF14* vicinity disrupt a *trans*-network of genes regulated by *KLF14* leading to metabolic defects in humans. The mouse model with *Klf14* deletion display reduced HDL-cholesterol in males, yet contrary to the human data, no phenotypes are observed in the female mice. Besides, no evidence of T2D characteristics are found in these *Klf14* knockout mice. Investigating the *Klf14* associated regulatory networks and phenotypes in the mouse and human genomes revealed species specific *Klf14* transcriptional targets, which may potentially be the reason for somewhat different *Klf14* associated phenotypes observed in mice and humans. Nevertheless, some *Klf14* transcriptional targets amongst the two species are recognised to interact with each other, suggesting their involvement in the same functional pathway. Moreover, a conserved regulatory motif associated with the *Klf14* phenotypes is identified between mice and humans. The regulatory networks of *Klf14* in the two species also share *Srebf2* (mouse orthologue of *SREBF1*), an important TF for lipid and cholesterol homeostasis. This data suggests that the functional role of *Klf14* in mice may have diverged to be largely involved in the cholesterol metabolism.

A comparative epigenetic profiling of the *KLF14* locus shows that the human adipose-specific enhancer harbouring few of the T2D risk variants is likely to be functionally inactive in the mouse. The region homologous to the human enhancer do not display any DNaseI hypersensitive activity in any of the mouse tissues available in ENCODE, which indicates a low probability of open chromatin or regulatory function at this loci. This suggests that distal regulatory elements associated with *Klf14* have possibly migrated and diverged in function, which reflects in the phenotypes associated with *Klf14* in mice and humans. This chapter specifically focuses on the epigenetic profile of the *KLF14* locus between the mouse and human genome. However, it would be interesting to compare the *KLF14* locus across several intermediate species as it could provide a detailed map of the functional divergence associated with *KLF14* locus regulatory activity. For instance, it could provide insights into the changing activity and position of the *KLF14* enhancer during evolution. Therefore, future study should focus on analysing the *KLF14* enhancer activity across several mammalian species with high quality genome builds such as human (*Homo sapiens*), macaque (*Macaca mulatta*), marmoset (*Callithrix jacchus*), mouse (*Mus musculus*), rat (*Rattus norvegicus*), rabbit (*Oryctolagus cuniculus*), cow (*Bos taurus*), pig (*Sus scrofa*), dog (*Canis familiaris*), and cat (*Felis catus*). However, such an analysis would require regulatory activity data such as enhancer annotations and/or DNase-seq profiles, in multiple tissues (or in at least adipose related tissues) across several species, which is not available at the moment. Some examples of cross-species regulatory data publicly available at present include H3K27ac and H3K4me3 profiles in adult liver across 20

mammalian species (Villar et al., 2015); H3K27ac profile in embryonic limb across human, rhesus and mouse (Cotney et al., 2013); and H3K27ac and H3K4me2 profiles in embryonic cortex across human, rhesus and mouse (Reilly et al., 2015). However, there is no repository with standardised H3K27ac profiles across different mammalian species in adipose tissue, therefore the data mining and processing for this analysis would be extensive. In addition to performing a cross-species analysis, comparing the *KLF14* locus within the different sub-strains of mice could detect the presence of any strain-specific variation in the *KLF14* locus associated enhancer activity (Lilue et al., 2018). Again, such cross-strain cross-tissue regulatory data in mice is not available in standardised repositories such as ENCODE, and hence very limited. However, an example of cross-strain regulatory data available in mice is the H3K27ac, H3K4me2 and ATAC-seq profiles in bone marrow derived macrophages across five mouse inbred strains - C57BL/6J, BALB/cJ, NOD/ShiLtJ, PWK/PhJ and SPRET/EiJ (Link et al., 2018). In the future when regulatory data for more species, strains and tissues become available, we will be able to study the *KLF14* regulatory function in adipose tissue across several genomes.

Overall, this chapter shows that such epigenetic differences should be taken into account when using or producing mouse models, especially if one intends to generate mouse models with non-coding mutations to model disease-associated SNPs (DA-SNPs) from humans. Mouse models with point mutations at homologous positions to the human non-coding DA-SNPs, which harbour no enhancer activity in the mouse are unlikely to produce phenotypes similar to the observed human traits. Additionally, this study highlights the need for enhancer annotations in the mouse genome to aid in the study of mammalian regulatory mechanisms and for systematic comparisons of regions between the human and mouse genomes.

By utilising the TFBS patterns conserved across species, I identified potential *cis*-regulatory variants in the *KLF14* T2D locus that may modulate *KLF14* expression and may contribute to the disease. A further positional bias analysis uncovered forkhead TFs to preferentially bind near three of the *cis*-regulatory variants in humans. Though none of the FKHD binding sites directly overlapped the SNP positions, forkhead TF *FOXO1* which binds within ± 25 bp of rs4731702, is known to be involved in insulin secretion and glucose production, and therefore, could be potentially involved in the mechanistic mode of action of these variants to cause metabolic defects. The SNP rs4731702 which has been previously identified to correlate with *KLF14* expression, emerged as one of the top candidates in both PMCA and positional bias analysis, hence providing evidence for the usefulness of such methods.

The methodologies applied here have limitations. First, the PMCA and positional bias analysis involves searching for only known TFBSs. This makes it dependent on

the current knowledge of TFs and their known binding sites, which is not complete. Also, the number of TF PWMs in publicly available databases is almost four times less compared to the commercial TF databases like TRANSFAC (Wingender et al., 1996), which further constraints this analysis. Second, detecting TFBSs using known PWMs involves a high rate of false positives which could lead to identifying binding sites with no functional significance *in vivo* - known as the “futility theorem" (Wasserman and Sandelin, 2004). Third, conservation of a sequence does not always correspond with regulatory activity. Many exceptionally conserved sequences have been recognised to show no important functional role (Ahituv et al., 2007). Conversely, numerous enhancers with weak or no conservation across distant species have been experimentally confirmed to be functionally active (Friedli et al., 2010; Taher et al., 2011). Moreover, the parameters applied to detect the CRMs would fail if the regulatory sequences have largely diverged across the species. These limitations could be addressed by using ChIP-seq profiles to detect the binding of a particular TF. However, as ChIP-seq profiles are tissue-specific and each experiment profiles a single TF, there is limited data available at the moment. Additionally, there is room for improvement in the implementation of the PMCA method. For instance, alignment of the input sequences and identification of conserved PWMs could be enhanced by modifying existing refined tools like CONREAL (Berezikov et al., 2004) and ConSite (Adams et al., 2000). Identification of the CRMs could also be improved by incorporating tools which do not depend on multiple genome alignments and take into account largely diverged regions (Cai et al., 2010). The lack of TF PWM data could be compensated by adding epigenetic information to the analysis such as DNaseI hypersensitive and histone modification data, which could help in detecting functionally active TFBSs with greater accuracy (Schwessinger et al., 2017).

Chapter 3

Identification of regulatory elements in the mouse genome

In this chapter, I model histone modification data from ENCODE to identify and characterise potential enhancer domains in a diverse set of mouse tissues and cell-types. The results described in this chapter contributes towards the following article:

Sethi, S., I. E. Vorontsov, I. V. Kulakovskiy, S. Greenaway, J. Williams, V. J. Makeev, S. D. M. Brown, M. M. Simon, A.-M. Mallon (2019).“Deciphering the impact of enhancer architecture on gene function and mouse phenotypes”. Under review in *Cell Reports*.

3.1 Introduction

The previous chapter displayed how examining regulatory elements such as enhancers can help unveil insights into molecular and disease mechanisms. If we look beyond the *KLF14* locus, over 90% of the GWAS SNPs associated with human disorders occur within non-coding regions (Hindorff et al., 2009), of which ∼76% are identified either within DHSs or in high LD with a SNP within a DHS (Maurano et al., 2012). This observation restates the possible role of active regulatory elements in human diseases. Moreover, a recent study identified 64% of the non-coding DA-SNPs from GWASs to occur in regions marked by H3K27ac (Hnisz et al., 2013), suggesting that the majority of these regulatory regions harbouring DA-SNPs are active enhancers. However, only a small number of functional enhancers (out of tens of thousands of putative enhancers) have been associated to gene expression changes. Additionally, the functional consequence of most of the DA-SNPs occurring within the enhancer regions stays unexplained, indicating that our insights about enhancers and their contribution in

diseases still remain insufficient. Given their functional importance, it is essential to identify mechanisms underlying transcriptional control by enhancers.

In order to understand the mechanisms involved in transcriptional regulation by enhancers, we need to: (1) identify the location of potential enhancer regions in the genome; (2) identify what TFs bind there; and (3) recognise the genes they potentially regulate. However, identification of enhancers is challenging as they are spread across the non-coding part of the genome. As $\sim$98% of the human genome is non-coding, we would have to scan billions of base pairs to distinguish potential enhancers from non-functional regions which is similar to looking for a needle in a haystack. Moreover, enhancer function is independent of its orientation and location, and could be present upstream; downstream; or within their target gene. Some enhancers can regulate genes several Mbs away, while some have been identified to regulate multiple genes (Mohrs et al., 2001), hence making the enhancer discovery process more confounded.

With the rapid progression of next-generation sequencing technologies in the past decade, it has become possible to more accurately predict genome-wide enhancer activity. Large scale programs such as the ENCODE (ENCODE Project Consortium, 2012), the FANTOM5 (FANTOM Consortium et al., 2014) and the NIH Roadmap Epigenomics project (Bernstein et al., 2010) have generated an initial detailed exploration of active enhancer and promoter regions in a plethora of tissues and cell-types, forming a crucial data source for the study of regulatory regions. ChIP-seq analysis of histone modifications has been most widely used by these consortiums to catalogue potential enhancer and promoter regions in humans, however, data and knowledge in the mouse is relatively low. At the time of analysing, no systematically defined genome-wide enhancer annotations were available in mouse tissues and cell lines. However, in the recent years, dense clusters of active enhancers known as super-enhancers (SEs), have been identified and characterised in an assortment of cells and tissues (Hnisz et al., 2013; Whyte et al., 2013), but they are limited to relatively few tissue types in the mouse (Adam et al., 2015; Fang et al., 2015; Ohba et al., 2015; Siersbæk et al., 2014; Vahedi et al., 2015; Whyte et al., 2013). Although two SE databases (Khan and Zhang, 2016; Wei et al., 2016) have been produced which provide SE locations generated using a standard analysis pipeline on publicly available data, they lack characterisation and do not provide other essential data such as the location of typical-enhancers (TEs) and their individual enhancer elements.

SEs have been shown to regulate important genes and master regulators that define the cell state (Whyte et al., 2013). Furthermore, significant enrichment of DA-SNPs from GWASs occur in SEs of disease-relevant tissues compared to TEs (Farh et al., 2015; Hnisz et al., 2013; Parker et al., 2013), but no study has systematically compared the enrichment of these SNPs in human and mouse SE domains. Therefore, identification

of SEs and other enhancer domains in a diverse set of mouse tissues could help in further comprehension of the regulatory mechanisms underlying gene function and disease causation. Knowing the location of enhancers in the mouse genome is important as it can facilitate molecular experiments to investigate the regulatory mechanisms in disease models. Furthermore, the location of enhancer regions can be further analysed to identify the binding sites of TFs within them. Additionally, techniques like CRISPR-Cas9 can be utilised to create mouse models mimicking the disease-causing genetic variation occurring within the enhancers in humans.

The majority of human diseases are a consequence of distorted interactions between cell- and tissue-type specific functions (Lage et al., 2008). For decades, pathologists have examined tissues from the affected organs to identify the presence, cause and extent of a disease. The relationship between cellular organisation, tissue structure and disease is well established. The cells in a tissue are required to perform common functions important for sustaining the tissue, and also unique functions that define the cell identity. These common and cell- or tissue-specific functions are governed by regulatory elements, which regulate the gene expression patterns. Therefore, studying tissue-specific regulatory regions and their networks have been of keen interest as they could provide insights into the biology of cell-type specific lineages. For instance, tissue-specific enhancer regions could aid in the identification of lineage-specific TFs and biomarkers. Although these elements have been comprehensively identified, numerous questions still remain on the interpretation of their biological relevance, effect on gene expression and overall impact on disease causation.

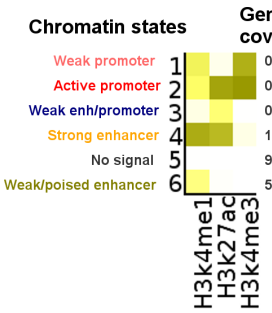
The aim of this chapter is to produce a catalogue of well defined enhancers in multiple mouse tissues and cell-types, which will allow us to further investigate the properties of these elements. In order to do this, I annotate potential regulatory elements in 22 mouse tissues using histone modification data. To investigate tissue-type specific biology, I systematically identify highly tissue-specific enhancer states across the tissues and further characterise them into SEs and TEs, henceforth producing a catalogue of multiple enhancer types in a diverse range of mouse tissues and cell-types, which includes previously unexplored tissues. Furthermore, I analyse open chromatin activity, TF binding occupancy and evolutionary sequence conservation within these newly identified enhancers. Finally, I investigate to what extent the occurrence of non-coding DA-SNPs from GWASs correlate between human and mouse enhancer domains. Overall, the results from this chapter provide valuable data and a platform to further characterise the mechanisms of gene regulation by enhancers in the mouse genome.

3.2 Results

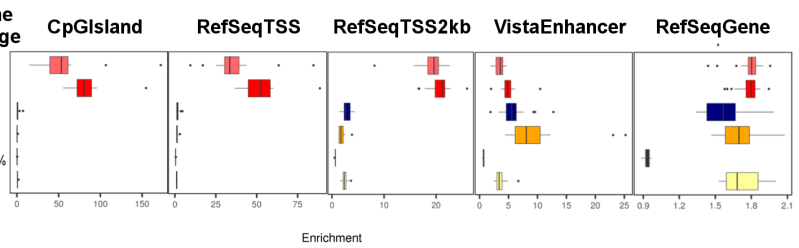
3.2.1 Chromatin state segmentation across 22 mouse tissues

To systematically identify genome-wide chromatin states and potential regulatory regions in the mouse genome, I applied a multivariate Hidden Markov model called ChromHMM (Ernst and Kellis, 2012) which models different patterns of histone marks observed in a dataset and assigns a state to each genomic position. Several ChromHMM models were produced consisting of 4-8 chromatin states (Appendix A.1) using ChIP-seq data of three primary histone marks (namely H3K4me1, H3K4me3 and H3K27ac) in 22 mouse tissues and cell lines. Visual inspection of these models identified the 6 state model to provide sufficient resolution in order to isolate chromatin states associated with promoters and enhancers. Using this model, I annotated the genome with 6 chromatin states (Fig. 3.1A) and assigned them likely biological roles based on the enrichment of histone marks (Fig. 3.1B). These chromatin states can be broadly categorised into active promoter, weak promoter, strong enhancer and weak enhancer states. To validate the predicted chromatin states, they were compared to promoters of known protein-coding genes, mouse enhancers from VISTA and other functional elements (such as known CpG islands, TSSs and gene coordinates) in the mouse genome. Active promoter and strong enhancer states achieved a recall sensitivity of 81.7% (18,543/22,707) with known promoters and 91.2% (331/363) with VISTA enhancers (described in section 1.6.3), respectively. As expected, chromatin states linked to promoters (state 1 and 2) overlap known CpG islands, TSSs and regions within 2 kb of TSSs (Fig. 3.1C). Conversely, chromatin states linked to enhancers (state 4 and 6) lack enrichment over TSSs and mostly overlap gene bodies and regions within 2 kb of TSSs. Examples of learned chromatin states are shown in Fig. 3.2. Overall, using ChromHMM, I annotated 427,251 weak promoter (state 1), 309,581 active promoter (state 2), 432,380 weak enhancer/promoter (state 3), 923,791 strong enhancer (state 4) and 2,531,993 weak enhancer (state 6) chromatin states across the 22 tissues, each chromatin state region being 200 bp in length and with a posterior probability > 0.95 in at least 1 tissue. It should be noted that these numbers represent the 200 bp resolution segmentations outputted by the ChromHMM, and not the actual number of predicted regulatory elements. To calculate the actual number of predicted regulatory elements, I concatenated adjacent chromatin state annotated regions together - this approach annotated on average 15,667 active promoter and 28,336 strong enhancer elements per tissue; or 54,564 non-redundant active promoter and 259,954 non-redundant strong enhancer elements across 22 tissues. These numbers are consistent with the mouse ENCODE study (Shen et al., 2012) which reported 53,834 putative non-redundant promoters and 234,764 potential enhancers across 19 mouse tissues.

A



C



B

State	Annotation	Enriched histone marks	Biological interpretation
1	**Weak promoter**	High enrichment of H3K4me3 and H3K4me1	H3K4me3 and H3K4me1 together are associated with promoters and TSSs. No enrichment of H3K27ac suggests these promoters might be weak or inactive.
2	**Active promoter**	High enrichment of H3K4me3 and H3K27ac; intermediate enrichment of H3K4me1	H3K4me3 and H3K4me1 together are associated with promoters and TSSs. H3K27ac mark is indicative of active regulatory elements.
3	**Weak enhancer/promoter**	Enrichment of H3K27ac	H3K27ac is a mark of active regulatory element and could be associated to promoters and enhancers.
4	**Strong enhancer**	High enrichment of H3K4me1 and H3K27ac	Enrichment of H3K4me1 combined with depletion of H3K4me3 is associated with enhancers. H3K27ac signal suggests these regions are active.
5	**No signal**	No enrichment	Depletion of all histone marks indicates no regulatory activity or low signal.
6	**Weak/poised enhancer**	Enrichment of H3K4me1	Enrichment of H3K4me1 combined with depletion of H3K4me3 is associated with enhancers and other distal elements. No enrichment of H3K27ac suggest these regions might be inactive or weak.

Fig. 3.1 Chromatin state segmentation and characterisation across 22 mouse tissues. (A) Heatmap showing the 6 state model emission parameters learned jointly across tissues by applying ChromHMM. Columns show the various histone marks used in the model, rows show the chromatin states with their potential functional annotation (interpreted from combinations of histone marks in each state), colour denotes the frequency with which a histone mark is observed in the genome, and genome-wide coverage of each state is shown on the right. (B) Table describing the histone mark signals enriched in each chromatin state and their functional interpretation. (C) Box plots showing the enrichment of each chromatin state over functional annotations across all the tissues (in order: CpG islands, TSSs, within 2 kb from TSSs, VISTA enhancers and genes). A higher enrichment value denotes a higher abundance of the chromatin state.

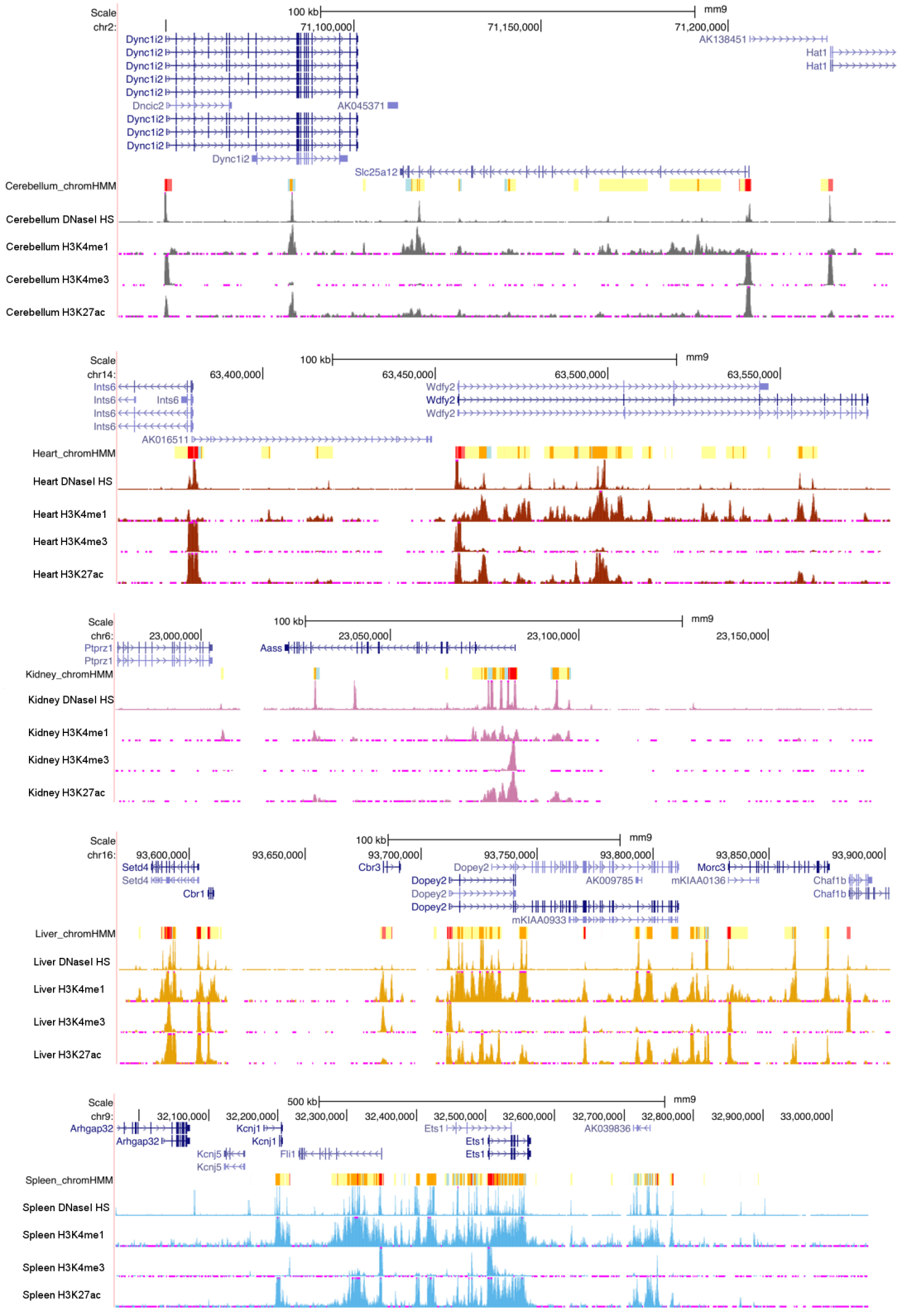


Fig. 3.2 Genomic view of chromatin state segmentation output. Genome browser snapshots displaying ENCODE ChIP-seq binding profiles of H3K4me1, H3K4me3 and H3K27ac; DNaseI hypersensitive (HS) signal; and this study's learned chromatin states. Chromatin states: ■ weak promoter; ■ active promoter; ■ weak enhancer/promoter; ■ strong enhancer; ■ weak/poised enhancer.

In order to validate the number of active promoter and strong enhancer elements predicted in this study, I compared them with two previously reported studies in mice: (1) study published by Shen et al. 2012 where they predicted enhancer, active promoter and insulator regions in 19 mouse tissues and cell types using genomic locations of H3K4me1, H3K27ac, H3K4me3, pol II binding and CTCF binding; (2) a study published by Bogu et al. 2015 where they predicted active promoters and enhancers by performing ChromHMM chromatin segmentation in 9 mouse tissues/cell lines using histone modifications (H3K4me1, H3K4me3, H3K27ac, H3K36me3, H3K27me3), CTCF binding and pol II binding. I found for most of the tissues, the number of predicted strong enhancer and active promoter elements are reduced in this study compared to previous studies (Fig. 3.3). This under-prediction may be due to: (1) a greater number of histone marks or other ChIP-seq data used by other studies in their prediction methods; (2) a strict threshold on posterior probability employed in this study (>0.95 as opposed to the default >0.75) to capture only the highly confident predictions. Compared to Shen et al. 2012 who reported to collectively annotate 11% of the mouse genome, this study assigned a potential regulatory function to 8.4% of the mouse genome.

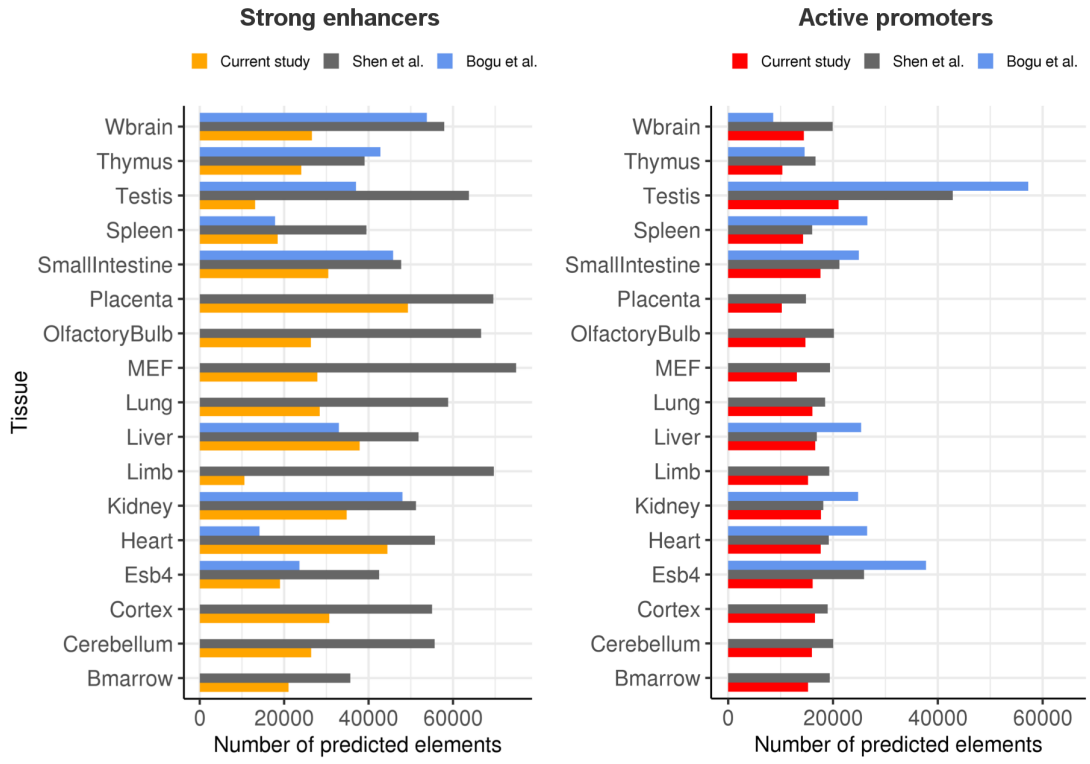


Fig. 3.3 Number of predicted regulatory elements in the mouse genome from three different studies. Comparison of the number of predicted regulatory elements in the mouse genome from Shen et al. 2012 (17 tissues), Bogu et al. 2015 (9 tissues) and the current study (22 tissues).

In order to compare the chromatin states between the different tissues, I computed the jaccard statistic which represents the similarity between two sets of genomic coordinates

based on their intersection. More specifically, the jaccard statistic represents the ratio of the intersection of two sets to the union of the two sets. Its value ranges between 0 and 1, where '0' represents no similarity between the two sets and '1' represents that the two sets are identical. Overall, the pairwise jaccard statistic correlations show that weak promoter and active promoter regions (state 1 and 2 respectively) are more similar between the tissues compared to strong enhancer and weak enhancer regions (state 4 and 6 respectively) (Fig. 3.4). As expected, one can also see that the chromatin states, especially active promoter and strong enhancer states, tend to have more overlap between physiologically similar tissues. For instance, functionally related tissues such as the cerebellum, cortex, whole brain and olfactory bulb cluster together in most chromatin states. Likewise, tissues and cell-types such as the spleen, thymus, bone marrow, MEL and CH12 group together. These observations support the functional relevance of these annotated enhancer and promoter states in gene expression and function. The number of annotations varies across different chromatin states, however, they are similar within the same chromatin state across majority of the tissues. The testis appears to be an exceptional case with a surprisingly high number of weak promoter annotations (state 1). Although it also contains large number active promoters (state 2), they are comparable to the numbers (for state 2) in other tissues. It remains undetermined as to why testis have a high number of promoter annotations. This could be an artefact or indicate the presence of noise in the data.

3.2.2 Open chromatin and TF binding activity

Chromatin organisation in DNA regulates TF binding activity by controlling the accessibility state of DNA. Regions of DNA which are accessible (or open chromatin) for TFs to bind are indicative of potential regulatory activity. To examine the relationship between the predicted regulatory elements and open chromatin regions, I compared the enhancer and promoter regions with DHSs in 11 mouse tissues (Fig. 3.5). On average, 93% of active promoters and 72% of strong enhancers overlap with DHSs across 11 tissues respectively. To examine if this overlap is more than what one would expect if these datasets were independent, a permutation test (with 1,000 permutations) was performed by shuffling the locations of enhancer and promoter regions to random locations in the mouse genome (see methods 3.3.2), which shows that strong enhancers and active promoters significantly overlap with DHSs ($p < 0.001$). A relatively small fraction of weak promoters (63%) and weak enhancers (30%) overlap DHSs (data not shown). This result further validates the predicted strong enhancer and active promoter annotations in the mouse tissues.

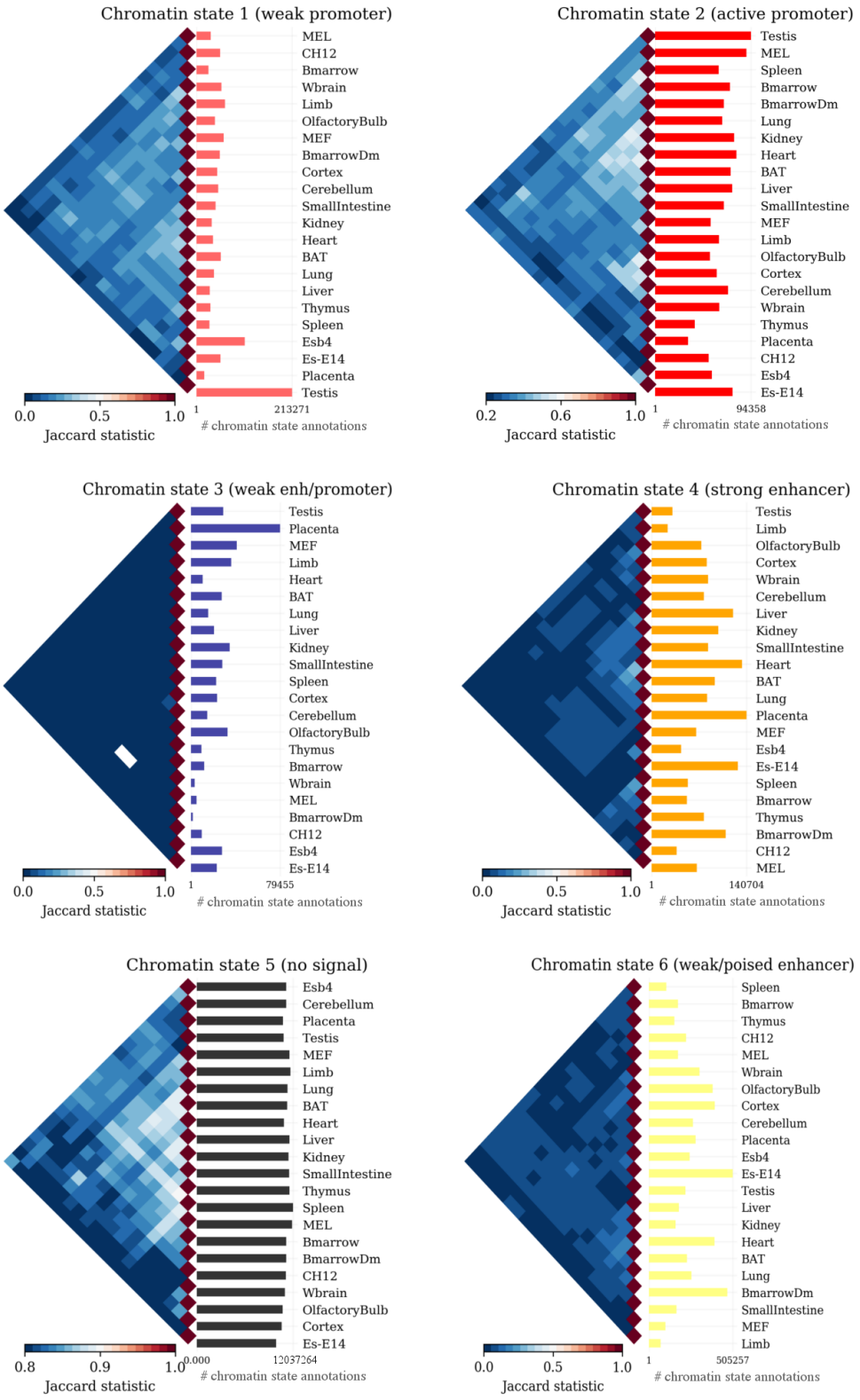


Fig. 3.4 Comparison of chromatin states across the tissues. Heatmaps display the pairwise jaccard statistic between the tissues in each chromatin state. A higher jaccard statistic denotes a higher two way intersection overlap of the chromatin state regions between the tissues. The bar plots on the right of the heatmaps display the number of chromatin state annotations in each tissue. BAT: brown adipose tissue; Bmarrrow: bone marrow; BmarrowDm: bone marrow derived macrophage; CH12: B-cell lymphoma; Esb4: mouse embryonic stem cells; Es-E14: mouse embryonic stem cell line at day E14.5; MEF: mouse embryonic fibroblast; MEL: leukaemia; Wbrain: whole brain.

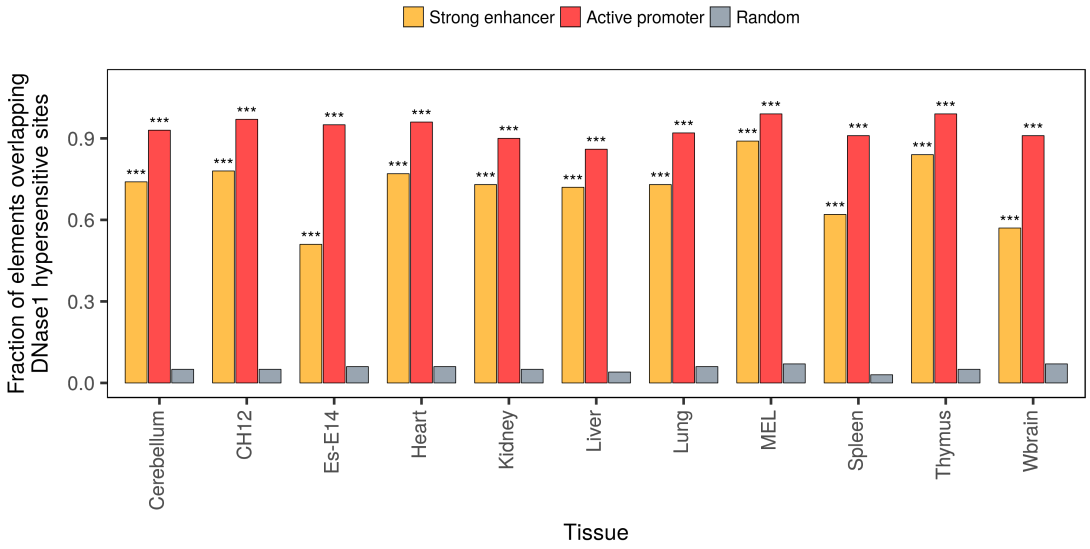


Fig. 3.5 Correlation between predicted regulatory elements and DHSs. Bar plot displaying the fraction of strong enhancers and active promoters overlapping DHSs in 11 mouse tissues. Statistical significance was computed by a permutation test. The permutation test was performed by shuffling the locations of enhancer and promoter regions to random locations in the mouse genome ('***' denotes $p < 0.001$). Grey bars display the median of the permuted distribution (random locations).

Next, I investigated the abundance of TF binding within enhancers and promoters. In order to do this, I compared the binding site locations of 71 different TFs (from 127 independent ChIP-seq datasets) to enhancer and promoter regions in the corresponding tissues. These 127 ChIP-seq datasets were manually extracted from open access databases and analysed to identify the TFBSs (Table 3.1) (see methods 3.3.2). Overall, approximately half of the TFs (60/127) have at least 30% of their binding sites overlapping strong enhancers and active promoters (permutation test, $p < 0.001$) (Fig. 3.6A-C). The fraction of binding sites correlating with strong enhancers and active promoters is variable between different TFs and possibly dependent on the biological function of the TF. Indeed, TFs have been previously shown to have a significant preference to bind either in promotors or distal regulatory elements (Cheng et al., 2014). Contrary to strong enhancers and active promoters, weak enhancers and weak promoters only harbour $\sim$13% of the TFBSs (Fig. 3.6D). These observations confirm that strong enhancer and active promoter regions are occupied with TFBSs.

Table 3.1 List of TFs analysed and the source of their ChIP-seq data.

ID	TF	Tissue	PMID / Project	ID	TF	Tissue	PMID / Project	ID	TF	Tissue	PMID / Project
TF1	Ar	Kidney	24451200	TF44	Jun	Spleen	22992523	TF87	Rad21	CH12	ENCODE
TF2	Batf	Spleen	22992523	TF45	Junb	Spleen	22992523	TF88	Rad21	MEL	ENCODE
TF3	Bhlhe40	CH12	ENCODE	TF46	Jund	CH12	ENCODE	TF89	Rcor1	CH12	ENCODE
TF4	Bhlhe40	MEL	ENCODE	TF47	Jund	MEL	ENCODE	TF90	Rcor1	MEL	ENCODE
TF5	Chd1	CH12	ENCODE	TF48	Jund	Spleen	22992523	TF91	Rest	Es-E14	22297846
TF6	Chd1	MEL	ENCODE	TF49	Kat2a	CH12	ENCODE	TF92	Runx1	Thymus	22325351
TF7	Chd2	CH12	ENCODE	TF50	Kat2a	MEL	ENCODE	TF93	Runx3	Spleen	24421391
TF8	Chd2	MEL	ENCODE	TF51	Klf4	Es-E14	18555785	TF94	Rxra	Liver	22158963
TF9	CTCF	Brain	ENCODE	TF52	Mafk	CH12	ENCODE	TF95	Sin3a	CH12	ENCODE
TF10	CTCF	Cbellum	ENCODE	TF53	Mafk	Es-E14	ENCODE	TF96	Sin3a	MEL	ENCODE
TF11	CTCF	CH12	ENCODE	TF54	Mafk	MEL	ENCODE	TF97	Six2	Kidney	22902740
TF12	CTCF	Es-E14	ENCODE	TF55	Max	CH12	ENCODE	TF98	Smc3	CH12	ENCODE
TF13	CTCF	Heart	ENCODE	TF56	Max	MEL	ENCODE	TF99	Smc3	MEL	ENCODE
TF14	CTCF	Kidney	ENCODE	TF57	Maz	CH12	ENCODE	TF100	Spl	Es-E14	24850855
TF15	CTCF	Limb	ENCODE	TF58	Maz	MEL	ENCODE	TF101	Spi1	MEL	22357756
TF16	CTCF	Liver	ENCODE	TF59	Mxi1	CH12	ENCODE	TF102	Stat3	Es-E14	18555785
TF17	CTCF	Lung	ENCODE	TF60	Mxi1	MEL	ENCODE	TF103	Stat5b	Liver	22158971
TF18	CTCF	MEL	ENCODE	TF61	Myb	MEL	ENCODE	TF104	Tal1	MEL	25409824
TF19	CTCF	Spleen	ENCODE	TF62	Myc	CH12	ENCODE	TF105	Tbp	CH12	ENCODE
TF20	CTCF	Thymus	ENCODE	TF63	Myc	MEL	ENCODE	TF106	Tbp	MEL	ENCODE
TF21	CTCF	Es-E14	18555785	TF64	Nanog	Es-E14	18555785	TF107	Tbx3	Heart	22706305
TF22	Egr2	Thymus	22306690	TF65	Nelfe	CH12	ENCODE	TF108	Tcf3	Spleen	20543837
TF23	Ep300	CH12	ENCODE	TF66	Nelfe	MEL	ENCODE	TF109	Tfcp2l1	Es-E14	18555785
TF24	Ep300	Es-E14	ENCODE	TF67	Nfe2	Liver	25128499	TF110	Tp53	Es-E14	22387025
TF25	Ep300	Heart	ENCODE	TF68	Nfya	Brain	22474351	TF111	Untf	CH12	ENCODE
TF26	Ep300	MEL	ENCODE	TF69	Nkx2-5	Heart	22706305	TF112	Untf	MEL	ENCODE
TF27	Esrrb	Es-E14	18555785	TF70	Nr5a2	Es-E14	20096661	TF113	Usf2	CH12	ENCODE
TF28	Ets1	CH12	ENCODE	TF71	Otx2	Es-E14	24905168	TF114	Usf2	MEL	ENCODE
TF29	Ets1	MEL	ENCODE	TF72	Pol2ra	Brain	ENCODE	TF115	Yy1	Es-E14	22210892
TF30	Foxa1	Liver	22737085	TF73	Pol2ra	Cbellum	ENCODE	TF116	Zc3h11	CH12	ENCODE
TF31	Foxa2	Liver	21623366	TF74	Pol2ra	CH12	ENCODE	TF117	Zc3h11	Es-E14	ENCODE
TF32	Gabpa	CH12	ENCODE	TF75	Pol2ra	Heart	ENCODE	TF118	Zc3h11	MEL	ENCODE
TF33	Gabpa	MEL	ENCODE	TF76	Pol2ra	Kidney	ENCODE	TF119	Zfx	Es-E14	18555785
TF34	Gata1	MEL	ENCODE	TF77	Pol2ra	Limb	ENCODE	TF120	Zkscan1	CH12	ENCODE
TF35	Gata4	Heart	25249388	TF78	Pol2ra	Liver	ENCODE	TF121	Zkscan1	MEL	ENCODE
TF36	Hcfc1	CH12	ENCODE	TF79	Pol2ra	Lung	ENCODE	TF122	Zmiz1	CH12	ENCODE
TF37	Hcfc1	Es-E14	ENCODE	TF80	Pol2ra	MEL	ENCODE	TF123	Zmiz1	MEL	ENCODE
TF38	Hcfc1	MEL	ENCODE	TF81	Pol2ra	Spleen	ENCODE	TF124	Znf143	Es-E14	23408857
TF39	Hnf4a	Kidney	24451200	TF82	Pol2ra	Thymus	ENCODE	TF125	Znf384	CH12	ENCODE
TF40	Hoxb4	Es-E14	22438249	TF83	Pou5f1	Es-E14	24905168	TF126	Znf384	Es-E14	ENCODE
TF41	Ikzf3	Thymus	22080921	TF84	Ppara	Liver	22158963	TF127	Znf384	MEL	ENCODE
TF42	Irf4	Spleen	22983707	TF85	Prdm4	Es-E14	23918801				
TF43	Jun	CH12	ENCODE	TF86	Prep1	Es-E14	25875616				

TF: transcription factor; PMID: Pubmed ID

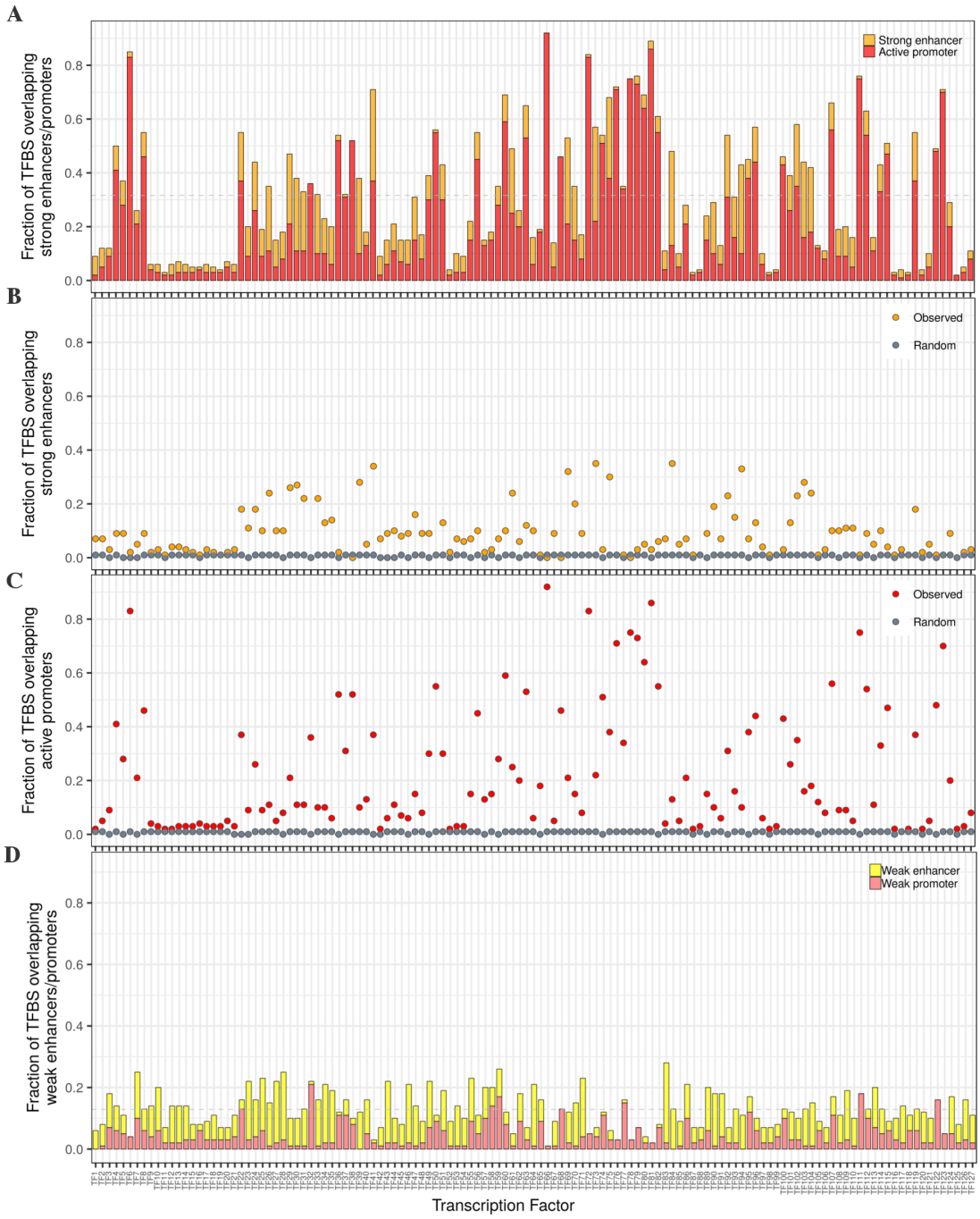


Fig. 3.6 Enrichment of TFBSs within regulatory elements. (A) Bar plot displaying the fraction of TFBSs within strong enhancers and active promoters. (B-C) Comparison between fraction of observed and randomly generated TFBSs within (B) strong enhancers and (C) active promoters. Grey data points denote the median of the permuted distribution for each TF. (D) Bar plot displaying the fraction of TFBSs within weak enhancers and weak promoters. The x-axis shows the ID of the TF from Table 3.1.

3.2.3 Identification of tissue-specific regulatory elements

Using the chromatin state probability outputted by ChromHMM for every 200 bp window in the genome, I constructed chromatin state posterior probability matrices of strong enhancer (consisting of 923,791 chromatin state annotations) and active promoter (consisting of 309,581 chromatin state annotations) such that each annotation had a posterior probability > 0.95 in at least one tissue. To identify tissue-specific regulatory elements (TSREs), I initially implemented k-means clustering to group elements with similar activity profiles across multiple tissues. Using this unsupervised clustering approach, strong enhancers and active promoters were grouped on the basis of their genomic locations across different tissues, revealing common and distinct clusters between the tissues. The strong enhancers and active promoters were grouped into 27 and 23 distinct clusters respectively (Fig. 3.7A-B).

Through visual inspection of the clusters, one can see that clusters 1-23 in strong enhancers and 1-14 in active promoters show a high to intermediate degree of tissue-specificity i.e. active in only one or few tissues. Whereas, clusters 24-27 in strong enhancers and 15-23 in active promoters display activity in multiple tissues. To quantify the amount of tissue-specificity amongst these cluster sets, the proportion of enhancers and promoters with respect to their activity in tissues was calculated (Fig. 3.7C-D). Enhancer cluster set 1-23 consisting of 843,239 chromatin state annotations (91% of total strong enhancers) are active in approximately three tissues on average (Fig. 3.7C), of which 491,525 (58% of chromatin state annotations in cluster set 1-23 or 53% of total strong enhancers) are specific to a single tissue. Likewise, a similar pattern is observed for promoter cluster set 1-14 consisting of 224,708 chromatin state annotations (73% of total active promoters), of which 140,587 (62% of chromatin state annotations in cluster set 1-14 or 45% of total active promoters) are active in only one tissue (Fig. 3.7D). These cluster sets display groups of enhancers and promoters common between tissues of organs which have similar functions or which work together. For instance, common enhancer regions were grouped together in immunity related tissues (clusters 4, 20, 22); brain tissues (clusters 6, 16, 23); placenta and ESCs (cluster 18); kidney and liver (cluster 10); and heart and lung (cluster 13). Interestingly, some enhancers and promoters are shared between BAT and heart (cluster 1) revealing a novel group of tissues which could possibly be connected. Recent studies have expanded the function of BAT in cardiovascular risk factors (such as glucose and lipid metabolism) and heart failure (Panagia et al., 2016; Thoonen et al., 2015; Thoonen et al., 2016). These enhancers which are active specifically in BAT and heart could possibly play a role in regulation of mechanisms associated with the cardiovascular system.

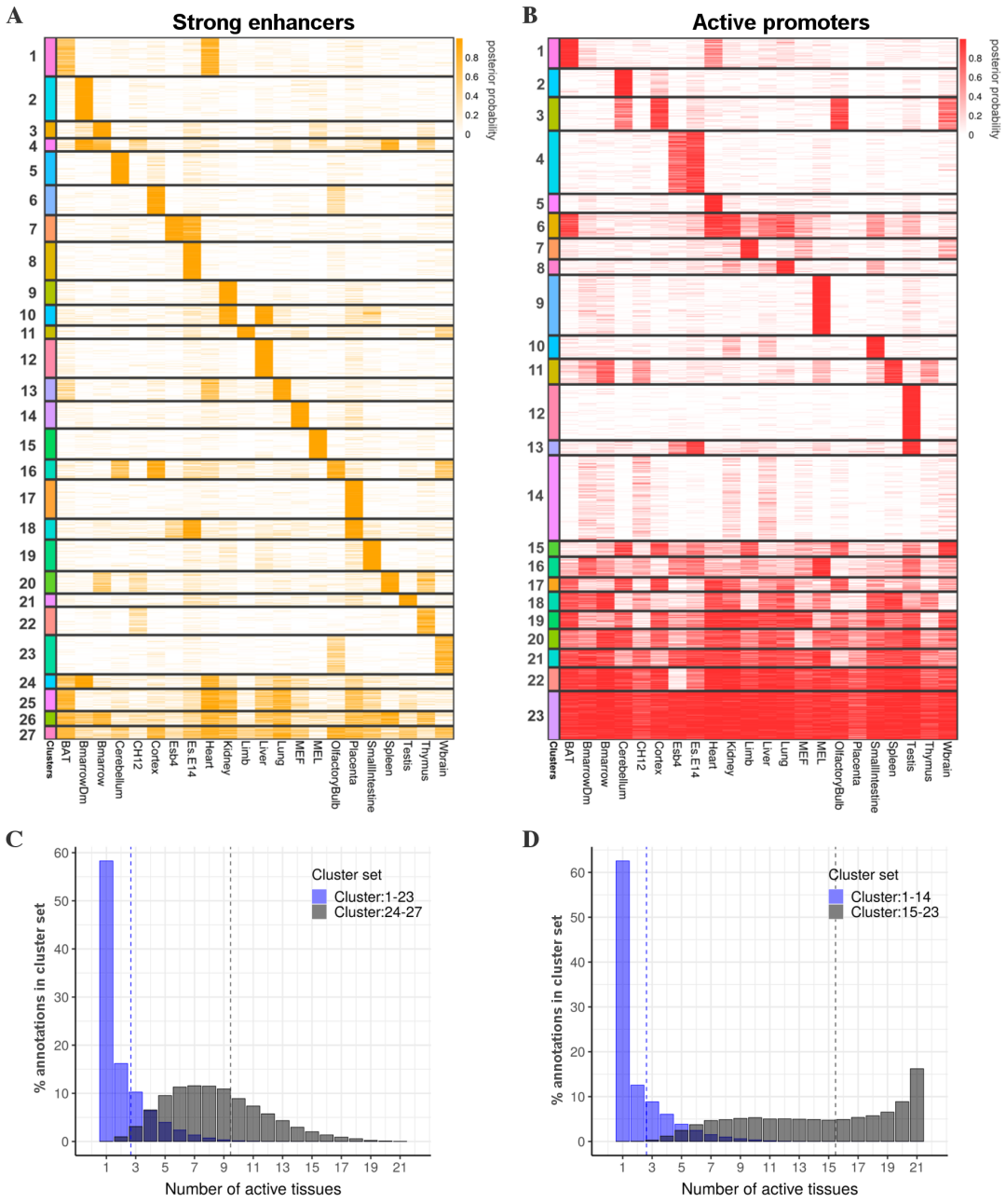


Fig. 3.7 Clustering of strong enhancers and active promoters across 22 mouse tissues. (A-B) Heatmaps displaying groups of (A) strong enhancers and (B) active promoters obtained by clustering their genomic locations across 22 tissues using k-means. Each row in the heatmap represents the location of an enhancer/promoter chromatin state and columns shows its corresponding posterior probability across different tissues. (C-D) Distribution of chromatin state annotations in (C) enhancer and (D) promoter cluster sets.

On the other hand, enhancer cluster set 24-27 consisting of 80,552 chromatin state annotations (9% of total strong enhancers) are not tissue-specific and are active in nine tissues on average (Fig. 3.7C). This suggests that only a small proportion of the enhancer landscape is similar across physiological different tissues. Compared to enhancers, a larger proportion (27%) of active promoters (cluster set 15-23 with 84,873 chromatin state annotations) is shared across the tissues, with activity in 15 tissues on average (Fig. 3.7D). This finding is consistent with previous studies (Ernst et al., 2011; Guenther et al., 2007b; Heintzman et al., 2009) where promoter groups were observed to be active across multiple cell lines and tissues. Overall, applying clustering to strong enhancer and active promoter regions identified moderately TSREs shared across $\sim$3 tissues on average.

In order to more accurately identify and classify TSREs, I implemented a previously described Tau algorithm (Kryuchkova-Mostacci and Robinson-Rechavi, 2017; Yanai et al., 2005) to calculate the tissue-specificity index of every regulatory element. The Tau algorithm has been utilised in past studies to quantify tissue-specific expression. The Tau method computes a score representing the tissue-specificity index ranging between 0 and 1, where a score of '0' indicates no tissue-specificity, while '1' indicates absolute tissue-specificity (see methods 3.3.4). Using the previously recommended thresholds of Tau score (represented as τ_{reg} for a regulatory element), I categorised enhancers and promoters into low ($\tau_{reg} \leq 0.15$), intermediate ($0.15 < \tau_{reg} < 0.85$), high ($0.85 \leq \tau_{reg} < 1$) and absolute tissue-specific ($\tau_{reg} = 1$). Chromatin state annotations with $\tau_{reg} \geq 0.85$ are observed to be highly tissue-specific, with the majority of annotations specific to a single tissue (strong enhancers: 96%, active promoters: 88%) (Fig. 3.8A-B). Whereas, chromatin state annotations with $\tau_{reg} \leq 0.15$ are shared across the tissues, with activity in 18 and 21 tissues on average for strong enhancers and active promoters respectively. This shows that the classification based on the Tau score is more specific compared to clustering. However, a portion of chromatin state annotations categorised as intermediate tissue-specific were active only in one tissue (strong enhancers: 32%, active promoters: 9%), possibly due to the strict thresholds used for τ_{reg} score. Lowering the τ_{reg} threshold used for highly tissue-specific group would help to recover some of the tissue-specific chromatin states, but it would also include noise i.e. chromatin states active in more than one tissue, therefore, the strict thresholds on τ_{reg} score were retained.

In total, I identified 284,677 ($\sim$31%) and 134,188 ($\sim$43%) highly tissue-specific ($\tau_{reg} \geq 0.85$) strong enhancer and active promoter chromatin states respectively across all the tissues, each region being 200 bp in length (Fig. 3.8C). Moreover, 3% of active promoters are shared across almost all the 22 tissues ($\tau_{reg} \leq 0.15$), while only 0.15% of strong enhancers are shared, suggesting enhancers to be variable across different tissues

and cell-types. Additionally, the Tau classification was implemented to weak promoter and weak enhancer states. Interestingly, ∼39% of weak promoter annotations are also identified to be highly tissue-specific. On the contrary, weak enhancer annotations are not tissue-specific, with only 2% as highly tissue-specific and the majority (97%) of them falling under the category of intermediate tissue-specificity. Since open chromatin regions can indicate potential active regulatory activity, I compared the captured highly tissue-specific strong enhancers and active promoter regions to DHSs in order to validate them. Both, tissue-specific strong enhancers and active promoters have a high degree of positive correlation with DNaseI signal in the corresponding tissues (Pearson's correlation, $p < 2.2 \times 10^{-16}$), confirming these mammalian TSREs to be highly tissue-specific (Fig. 3.9). These TSREs identified by the Tau method were used for all the subsequent analysis.

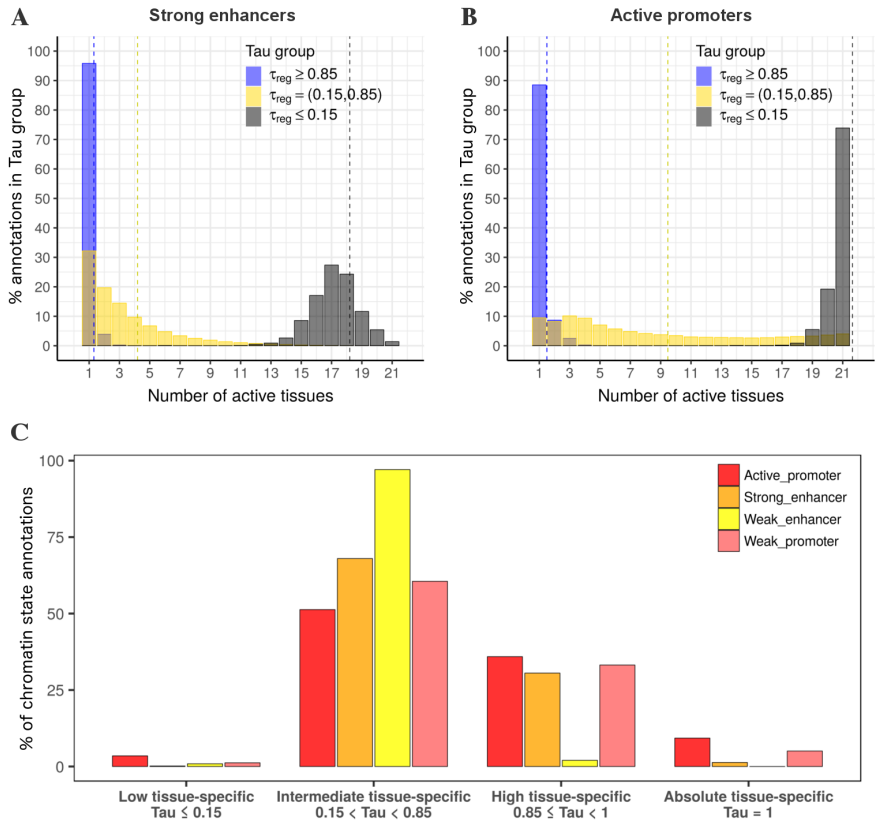


Fig. 3.8 Distribution of tissue-specific regulatory elements. (A-B) Distribution of chromatin state annotations in (A) enhancer and (B) promoter groups categorised according to their tissue-specificity index, as measured by τ_{reg}. (C) Percentage of chromatin state annotations in different chromatin states categorised according to their tissue-specificity index.

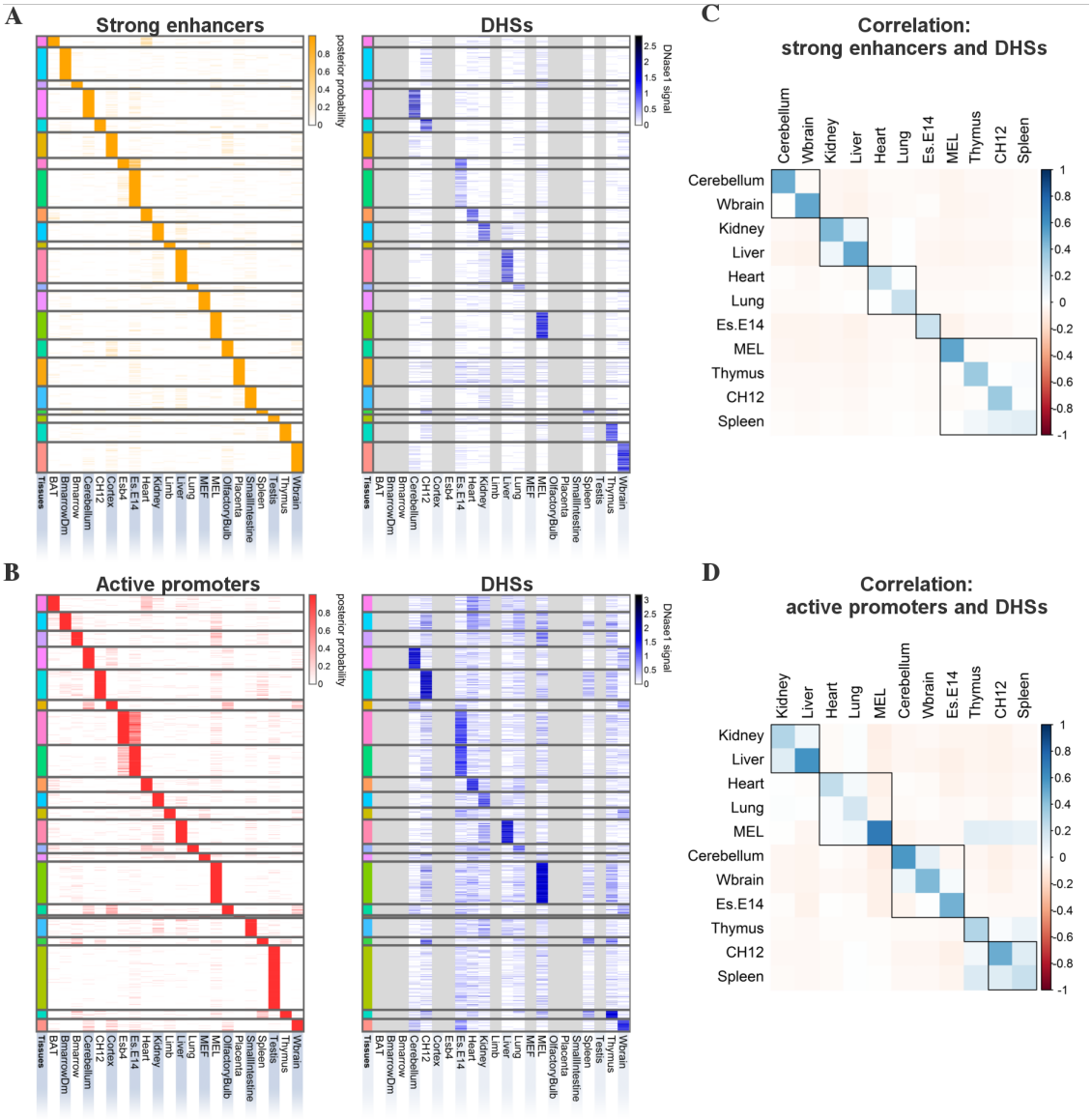


Fig. 3.9 Tissue-specific regulatory elements in 22 mouse tissues. (A) Strong enhancers (B) Active promoters. Heatmaps showing chromatin state posterior probability of tissue-specific chromatin states ($\tau_{reg} \geq 0.85$) (left) and their corresponding DNAseI signal (right) in every tissue. Each row is a genomic location and columns represent different mouse tissues and cell lines. Grey columns show tissues for which data was not available. The heatmaps have been sorted by the order of the tissues across the columns. (C-D) Heatmaps showing pairwise Pearson's correlations between (C) tissue-specific strong enhancers and DHSs; and (D) tissue-specific active promoters and DHSs. The order of the tissues is sorted by hierarchical clustering and the boxes represents clusters obtained from the clustering of the correlation matrix.

3.2.4 Detection of super-enhancers in the mouse genome

To detect SEs in the mouse genome, the ROSE algorithm (Whyte et al., 2013) was used to stitch together tissue-specific enhancers within a span of 12.5 kb into cohesive units and rank them based on H3K27ac signal (Fig. 3.10A). The stitched enhancers with high H3K27ac enrichment are defined as SEs and are systematically segregated from the TEs. The tissue-specific enhancers form the building blocks of these cohesive units (for both categorised as SEs or TEs) and are referred to as constituent enhancers (Fig. 3.10A). Using this approach, 6.6% (5,082) of all cohesive units (or 24% of all tissue-specific enhancers) were identified as SEs while 93.4% (71,824) are TEs (76% of all tissue-specific enhancers), hence generating a comprehensive catalogue of SEs and TEs in 22 mouse tissues (Appendix A.2). Consistent with previous research, SE cohesive units are occupied on average by 2.4× H3K27ac and span large genomic regions (median size = 12.4 kb) compared to TEs (median size = 0.4 kb) (Appendix A.3). Enrichment of H3K4me1 and DNaseI at SEs is observed to be in agreement with H3K27ac levels (Appendix A.4). High enrichment of these histone marks and pol II is also detected at constituent enhancers within the SEs compared to TEs (2× higher H3K27ac; 1.3× higher H3K4me1; 1.4× higher pol II), suggesting increased levels of chromatin activity in SEs (Appendix A.5). As an example, the segregation of SEs in cerebellum is shown in Fig. 3.10B-E.

3.2.5 Evolutionary conservation of mouse enhancers

Prior investigations to understand enhancer evolution have indicated that enhancers and TF binding have diverged significantly across some mammals (Odom et al., 2007; Shibata et al., 2012; Yue et al., 2014). A recent study of enhancers across 20 mammalian species (Villar et al., 2015) demonstrated that enhancers have evolved at a faster rate compared to promoters and are rarely conserved across mammals. However, most enhancer regions associated with embryonic development are highly conserved due to highly similar key developmental programs across mammals (Pennacchio et al., 2006). For instance, highly conserved long stretches (> 200 bp) of DNA segments (referred to as'ultraconserved' regions) are preferentially located near TFs which are known to be involved in embryonic development, suggesting their potential role in controlling essential developmental genes (Bejerano et al., 2004; Dickel et al., 2018b; Sandelin et al., 2004).

In order to gain evolutionary insights about these newly identified SEs and TEs in the mouse, I investigated the sequence conservation of SE and TE constituent enhancers. The sequence conservation of enhancer regions was quantified using the

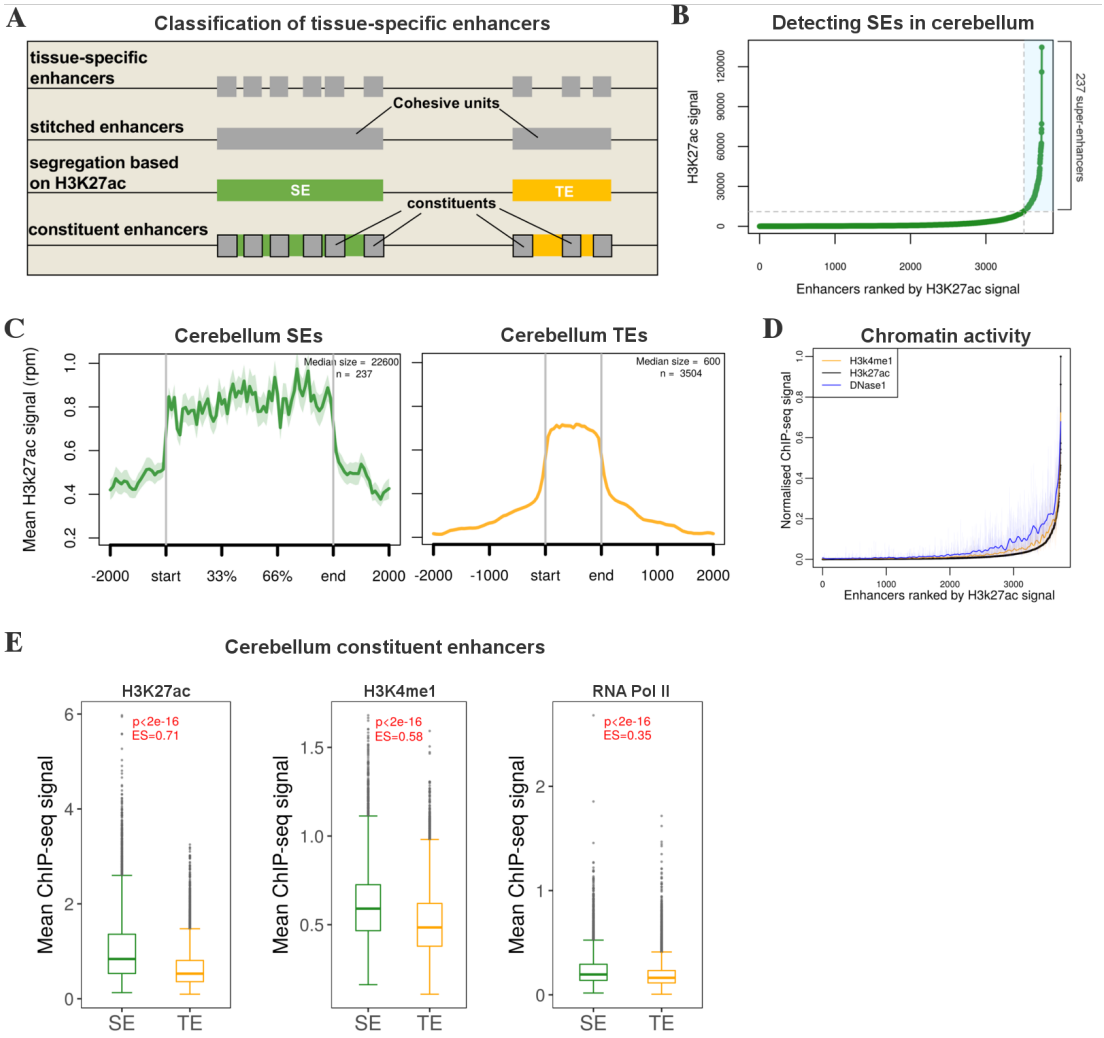


Fig. 3.10 Detection of SEs in the mouse genome. (A) A schematic illustrating the methodology used to classify tissue-specific enhancers into SEs and TEs. The tissue-specific enhancers are stitched together into cohesive units and ranked on their H3K27ac density. High ranking stitched enhancers are defined as SEs. The original tissue-specific enhancer elements within SEs and TEs are referred to as constituent enhancers. (B) Distribution of H3K27ac ChIP-seq signal over cerebellum-specific enhancers stitched together within 12.5 kb (n = 3,741). Stitched cohesive units (x-axis) are ranked in an increasing order of their input-normalised H3K27ac signal (reads per million, y-axis). This approach identified 237 SEs (highlighted in blue) and 3,504 TEs in the cerebellum. (C) Metagene profile of mean H3K27ac ChIP-seq signal across all the SEs and TEs in cerebellum. The profiles are centred on the enhancer regions and the surrounding 2 kb regions around each enhancer is shown. The length of the enhancer region is scaled to represent the median size of SEs (22,600 bp) and TEs (600 bp) in cerebellum. The shaded area shows the standard error (SEM). See also Appendix A.3. (D) Distribution of H3K4me1, H3K27ac and DNaseI signal across stitched enhancers in the cerebellum. The stitched enhancers for each feature on the x-axis are ranked according to the H3K27ac ChIP-seq signal. See also Appendix A.4. (E) Comparison of H3K27ac, H3K4me1 and pol II ChIP-seq signal between SE and TE constituent enhancers in the cerebellum. See also Appendix A.5. p: p-values from Wilcoxon Rank Sum Test; ES: non-parametric effect size (see methods 3.3.6).

phastCons score (Siepel et al., 2005) calculated from alignments of 20 mammalian species (see methods 3.3.7). The average phastCons score of enhancers was compared to a background set of random size-matched genomic regions (control regions). It is observed that SEs either have equal ($p > 0.05$, $ES = 0$), or significantly lower ($p < 10^{-3}$, $ES \leq -0.43$) conservation scores compared to control regions in adult tissues, though the magnitude of difference in the majority of the tissues is small ($ES < -0.20$) (Fig. 3.11). This suggests that SEs generally lack sequence conservation across mammalian species. However, in embryonic tissues (such as Esb4, Es-E14, placenta and whole brain), SEs exhibit significantly higher sequence conservation compared to control regions ($p < 10^{-4}$, $ES \leq 0.26$). Likewise, TEs exhibit either equal ($p > 0.05$, $ES = 0$), or lower ($p < 10^{-2}$, $ES < -0.20$) sequence conservation relative to control regions in adult tissues, with the exception in bone marrow derived macrophages, liver, MEL and thymus, where TEs have significantly higher conservation scores ($p < 10^{-4}$, $ES = 0.15$) (Fig. 3.11). Similar to SEs, TEs are also highly conserved compared to control regions in embryonic tissues such as Esb4, Es-E14, limb and whole brain ($p < 10^{-4}$, $ES \leq 0.53$). This result substantiates with previous research showing enhancers in embryonic tissues to be more conserved and possibly under a stronger evolutionary constraint compared to enhancers in adult tissues (He et al., 2011; Nord et al., 2013).

Next, I compared the sequence conservation between SE and TE constituent enhancers. Compared to TEs, SE constituents in majority of the tissues (13/22) have significantly lower sequence conservation scores ($p < 10^{-2}$, $ES \leq -0.49$), indicating that TEs are generally more conserved than SEs. Whereas for the remaining tissues, the sequence conservation scores are not different between SE and TE constituents ($p < 0.05$), with 7 out of 9 tissues achieving an effect size of 0. Overall, SE and TE constituents appear to be poorly conserved across the 20 mammalian species analysed here, with SEs showing substantially lower sequence conservation. A previous study comparing functional enhancers (identified using H3K27ac) in the liver across 20 mammalian species (Villar et al., 2015) revealed that enhancers are rarely functionally conserved across mammals, suggesting that they evolved more recently. Since SEs have poor sequence conservation, it is possible that they evolved more recently, however it is difficult to confirm this based on only sequence conservation, as a lack of sequence conservation does not necessarily mean that the sequences are not functional (Cooper and Brown, 2008).

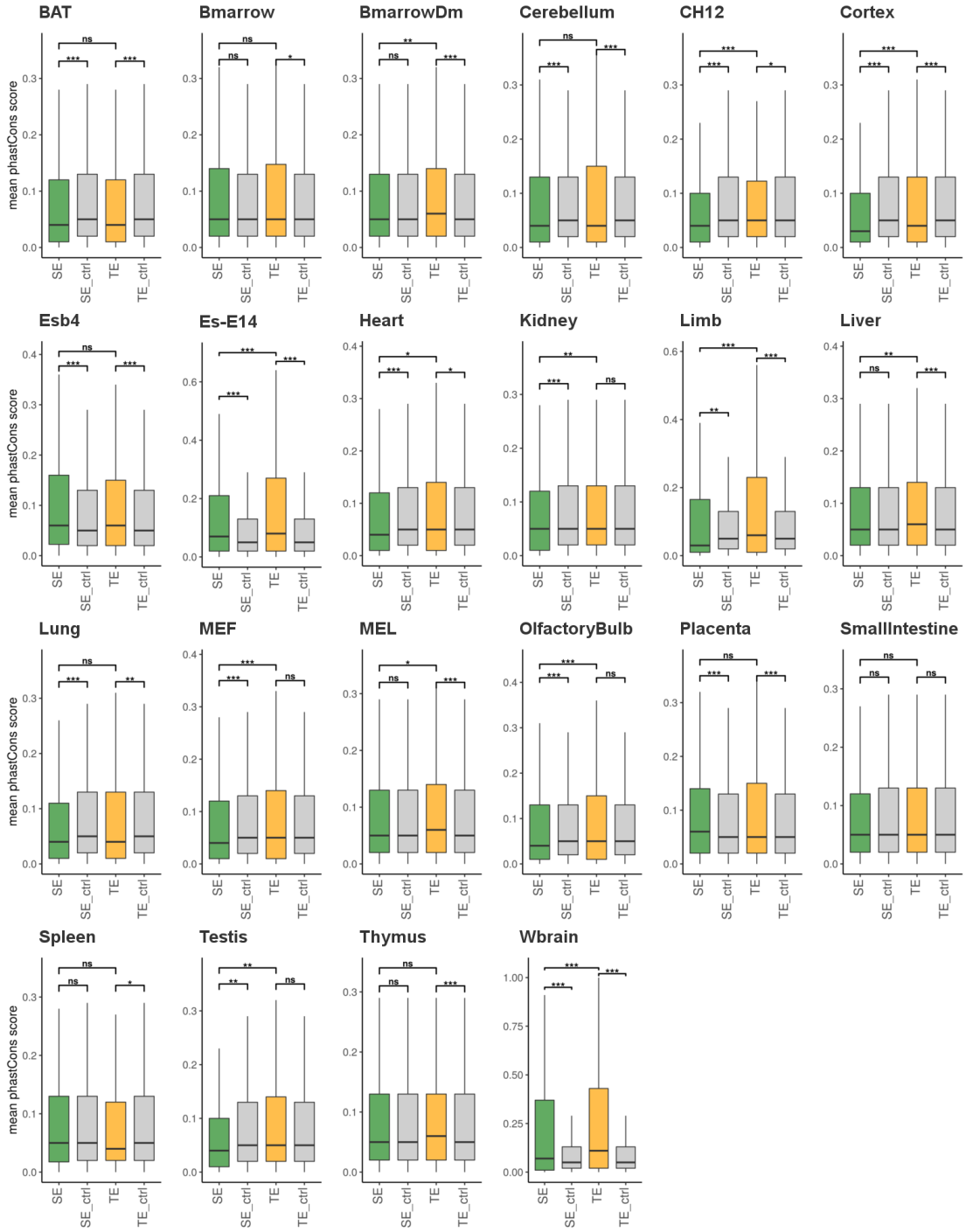


Fig. 3.11 Sequence conservation of mouse enhancers across 20 mammalian species. Box plots displaying the sequence conservation score (phastCons) of SE and TE constituent enhancers in each tissue. 'SE_ctrl' and 'TE_ctrl' represent the background size-matched control regions for SEs and TEs respectively. The control regions were generated by shuffling the enhancer locations to random genomic sites with identical size and number (1,000 permutations). P-values were calculated using the Wilcoxon Rank Sum Test ('***' denotes $p < 0.0001$, '**' denotes $p < 0.001$, '*' denotes $p < 0.01$, 'ns' denotes not significant).

3.2.6 Disease-associated SNPs in mouse enhancers

Recent studies have shown DA-SNPs from GWASs to be more enriched in SEs compared to TEs (Farh et al., 2015; Hnisz et al., 2013). Moreover, DA-SNPs tend to occur in SEs of disease-relevant tissues and cell-types. For instance, 19% (13/67) of non-coding SNPs associated with type 1 diabetes occur within SEs of primary T-helper cells as opposed to 9% (6/67) in TEs (Hnisz et al., 2013). Therefore, I sought to examine whether disease-associated genetic variation occurs in mouse enhancer regions of disease-relevant tissues, which could identify potential regions in the mouse for functional characterisation of DA-SNPs. For this purpose, I compared the enrichment of 1,592 non-coding DA-SNPs, comprising of 26 different phenotypic traits and diseases from GWASs, between SEs and TEs in the human and mouse genomes (Fig. 3.12) (see methods section 3.3.8 for details of SNP and trait selection). The human SEs and TEs were retrieved from Hnisz et al. (2013) in tissues and cell-types which were similar to the tissues in the mouse dataset for an unbiased comparison (n=15). The coordinates of human DA-SNPs were converted to mouse positions which resulted in 820 (∼51%) mouse SNPs homologous to the human DA-SNPs, while the remaining 49% of non-coding DA-SNPs did not have a homologous position in the mouse genome.

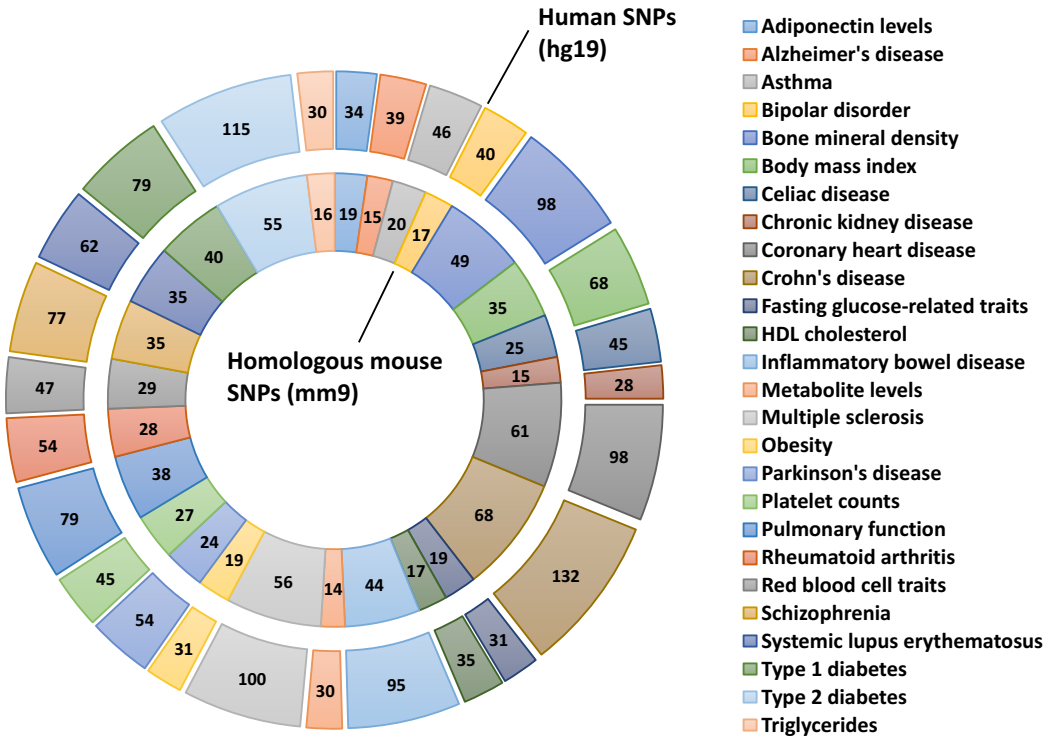


Fig. 3.12 Non-coding DA-SNPs associated with 26 phenotypic traits and diseases. Chart displaying the number of non-coding DA-SNPs associated with each of the GWAS trait analysed in this study.

To summarise the analysis, the comparison of each GWAS trait was placed into four categories: (1) category A, DA-SNPs enriched in enhancers of disease-relevant tissues in both human and mouse; (2) category B, DA-SNPs enriched in enhancers of disease-relevant tissues in only human; (3) category C, DA-SNPs enriched in enhancers of disease-relevant tissues in only mouse; and (4) category D, DA-SNPs not enriched in enhancers of disease-relevant tissues in both human and mouse. Of the 26 GWAS traits, 5 were placed in category A, 7 in category B, 4 in category C and 10 in category D (Table 3.2). Further inspection of trait enrichment in category C revealed that for 3 traits (out of 4), the disease-relevant tissue was not available in the human enhancer dataset which is possibly the cause for DA-SNPs not enriched in human enhancers. Another potential reason could be the different datasets and processing used to identify human enhancer regions by Hnisz et al. (2013), which could lead to differences in enhancer annotation. An example from each category is displayed in Fig 3.13 (see Appendix A.6 for all traits).

Table 3.2 Comparison of DA-SNPs enrichment in human and mouse enhancer regions.

Phenotypic trait/disease (GWAS)	Result category
Celiac disease	A
Multiple sclerosis	A
Rheumatoid arthritis	A
Systemic lupus erythematosus	A
Type 1 diabetes	A
Alzheimer's disease	B
Coronary heart disease	B
Crohn's disease	B
Inflammatory bowel disease	B
Platelet counts	B
Red blood cell traits	B
Schizophrenia	B
Asthma	C
Chronic kidney disease	C
Fasting glucose-related traits	C
HDL cholesterol	C
Adiponectin levels	D
Bipolar disorder	D
Bone mineral density	D
Body mass index	D
Metabolite levels	D
Obesity	D
Parkinson's disease	D
Pulmonary function	D
Type 2 diabetes	D
Triglycerides	D

Table displaying the comparative enrichment of DA-SNPs within human and mouse enhancers. Result category codes: A, DA-SNPs enriched in enhancers of disease-relevant tissues in both human and mouse; B, DA-SNPs enriched in enhancers of disease-relevant tissues in only human; C, DA-SNPs enriched in enhancers of disease-relevant tissues in only mouse; D, DA-SNPs not enriched in enhancers of disease-relevant tissues in both human and mouse.

Furthermore, confirming previous reports, DA-SNPs were more enriched in SEs of disease-relevant tissues relative to TEs in both the human and mouse genomes. Overall, excluding the GWAS traits in category D in which the DA-SNPs were not enriched in human enhancers, 9 out of 16 GWAS traits have SNPs enriched in mouse enhancers of disease-relevant tissues, suggesting potentially conserved regulatory pathways between the human and mouse genomes. Interestingly, all the GWAS diseases in category A are related to the immune system (Table 3.2), however, translating data from mouse models of immune diseases into human research has been largely unsuccessful, due to the differences in immune response mechanisms between the mouse and humans (Zschaler et al., 2014). Perhaps targeting the category-A DA-SNPs homologous positions in the mouse could produce disease models to better translate the disease understanding into human research. This emphasises the importance of taking the differences in the enhancer landscape between human and mouse into account when using mouse models to study gene regulation (as discussed in chapter two). Such analysis can help in predicting the DA-SNPs and phenotypic domains which could be better replicated in mouse models via genome editing techniques.

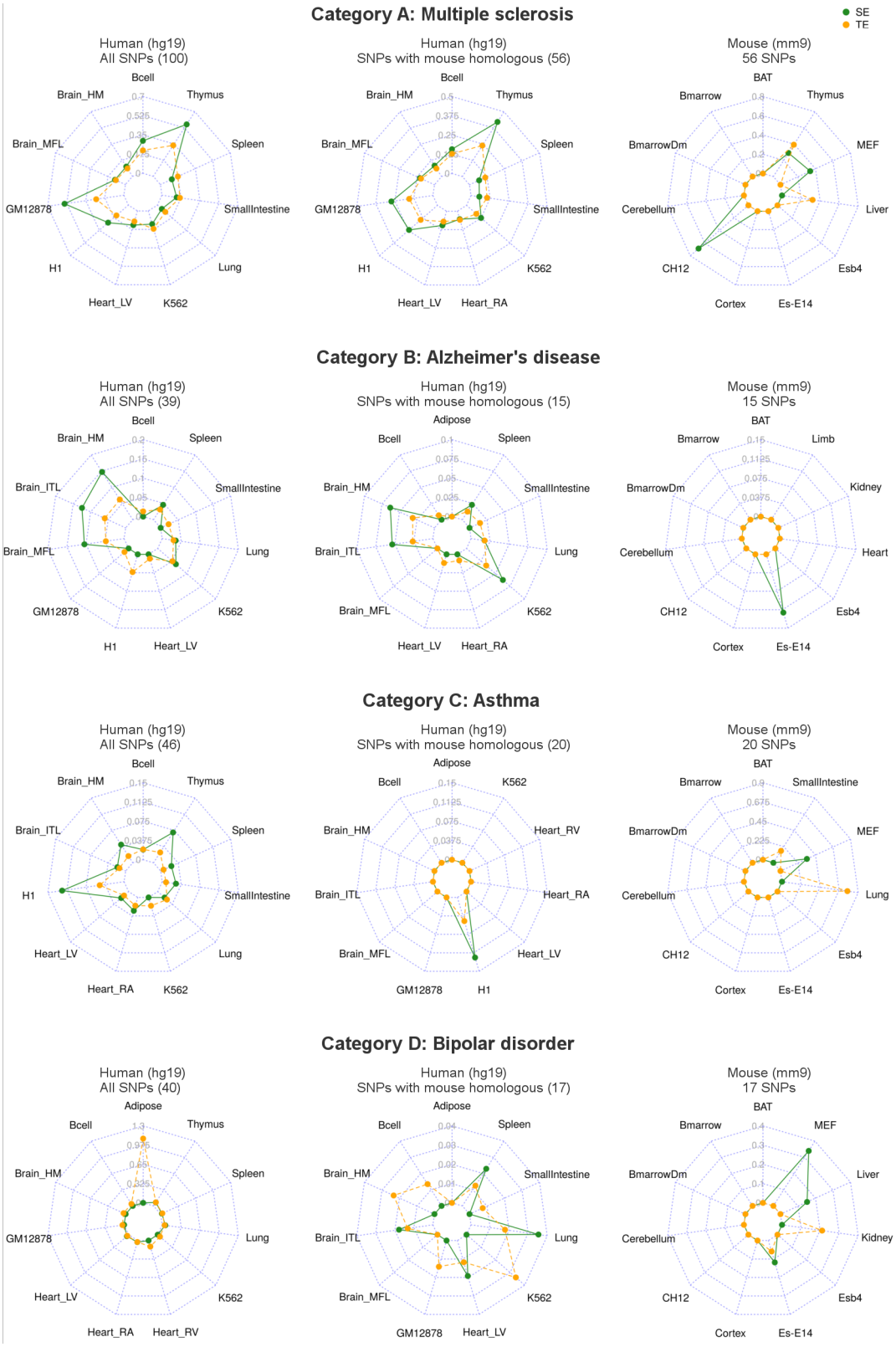


Fig. 3.13 Enrichment of disease-associated genetic variants in human and mouse enhancers. Radar plots displaying examples of DA-SNP enrichment in SEs and TEs of human and mouse. Each respective axis shows the SNP density (SNP per Mb of enhancer) in SE and TE domains (green and orange dots respectively). Brain_HM: Brain hippocampus middle; Brain_MFL: Brain mid frontal lobe; Brain_ITL: Brain inferior temporal lobe; Heart_LV: Heart left ventricle; Heart_RV: Heart right ventricle; Heart_RA: Heart right atrium.

3.3 Methods

3.3.1 Learning Chromatin states in the mouse genome

Firstly, the ChIP-seq data for histone H3 lysine 4 monomethylation (H3K4me1), histone H3 lysine 4 trimethylation (H3K4me3) and histone H3 lysine 27 monoacetylation (H3K27ac) in 22 mouse tissues was collected from the the ENCODE project (LICR lab) in the form of sequence alignments (BAM files mapped to mm9 mouse genome). The dataset includes 14 adult tissues: BAT (brown adipose tissue), bone marrow, cerebellum, cortex, heart, kidney, liver, lung, olfactory bulb, placenta, small intestine, spleen, testis and thymus; 2 embryonic tissues: limb and whole brain; and 6 cell lines: bone marrow derived macrophage, CH12 (B-cell lymphoma, GM12878 analogue), Esb4 (mouse embryonic stem cells), Es-E14 (mouse embryonic stem cell line at E14.5), MEF (mouse embryonic fibroblast), MEL (leukaemia, K562 analogue). Next, I used ChromHMM (a multivariate Hidden Markov model) to integrate all the ChIP-seq data and summarise into easily illustratable annotations. The chromatin states were jointly learned across 22 mouse tissues using default parameters. Several ChromHMM models were produced consisting of 4-8 chromatin states, of which the 6 state model was identified to provide sufficient resolution in order to isolate biologically meaningful chromatin states. The resulting chromatin states were then annotated based on the biological significance of the frequencies of combined histone marks (Fig. 3.1). Using this approach, potential active promoter (404,016), weak promoter (647,185), strong enhancer (1,075,608) and weak enhancer (2,068,844) chromatin state annotations were mapped across 22 mouse tissues and cell-types. To validate the predicted promoter states (states 1 and 2), I compared 217,678 unique non-overlapping predicted promoter chromatin states to 22,707 known protein-coding genes (mm9 ensembl genes v67; 10 kb upstream, 100 bp downstream of TSSs) and recovered 81.66% of known promoters. Similarly, to validate the strong enhancer predictions (state 4), I compared 386,222 unique non-overlapping enhancer chromatin states to 363 experimentally validated VISTA mouse enhancers and recovered 91.18% of VISTA enhancers. Chromatin states with $\leq$0.95 posterior probability were filtered resulting in 923,791 strong enhancer (state 4); 309,581 active promoter (state 2); 2,531,993 weak enhancer (state 6); and 427,251 weak promoter (state 1) high confidence chromatin state annotations respectively.

3.3.2 Comparing regulatory elements with DHSs and TFBSs

To compare strong enhancers and active promoters with open chromatin regions, DHSs in 11 epigenomes (Cerebellum, CH12, Es-E14, Heart, Kidney, Liver, Lung, MEL, Spleen, Thymus, Wbrain) were collected from the ENCODE (University of Washington

lab) in the form of hotspots. The fraction of enhancer/promoter regions overlapping DHSs was calculated as: number of base pairs in enhancers/promoters intersecting DHSs divided by the total enhancer/promoter coverage. In order to examine if this overlap was statistically significant compared to what one would expect if they were independent, a permutation test was performed by shuffling the enhancers and promoters to random locations in the genome (n=1,000). These random locations were compared to DHSs to build a background distribution and the p-value was calculated as the number of permuted overlap greater than the observed overlap, divided by the total number of simulations. In order to investigate TF binding within enhancers and promoters, 127 TF ChIP-seq datasets were manually extracted from different publicly available resources. The binding sites for 7 TFs were extracted from JASPAR, while the rest of the TF ChIP-seq peaks were extracted from the Cistrome browser (Mei et al., 2017) and the ENCODE project. The ChIP-seq peaks were then analysed using Homer to find the enriched motifs and the binding sites of the canonical motif associated with the TF. The overlap between TFBSs and enhancer/promoter elements along with permutation test was computed in a similar fashion as described above for the overlap with DHSs.

3.3.3 Clustering of promoters and enhancers across 22 tissues

Clustering was performed on active promoter (state 2) and strong enhancer (state 4) states. The complete mouse genome was segmented into 200 bp intervals and regions annotated as active promoter or strong enhancer states (posterior probability > 0.95) in at least one tissue were extracted to be used in clustering. Using this approach, a chromatin state posterior probability matrix was constructed for strong enhancers (state 4 posterior probability > 0.95) and active promoters (state 2 posterior probability > 0.95) with dimensions $n \times s$, where n is the number of chromatin state annotations and s is the number of tissues (i.e. 22). Each row of the matrix was a genomic location of the chromatin state and columns represented its posterior probability across all the tissues. The clustering was performed on the rows of this matrix using the k-means algorithm in R. By calculating the sum of squared distances of samples to the nearest cluster centre combined with visual inspection, I found 23 and 27 to be the optimal number of clusters into which the promoter and enhancer data could be divided to reveal distinct groups respectively. The clusters were then reordered for better visualisation in a heatmap.

3.3.4 Tissue-specificity index of regulatory elements

To identify tissue-specific regulatory regions across the 22 tissues, I implemented the Tau method which has been used to detect tissue-specific expression in previous studies

(Kryuchkova-Mostacci and Robinson-Rechavi, 2017; Yanai et al., 2005). Tau is a measure of tissue-specificity index which takes into account the number of tissues and normalised expression in each tissue, and outputs a score for each gene. The Tau method was applied to the posterior probability matrices constructed in section 3.3.3. The Tau score for each regulatory element was calculated by the following equation:

$$\tau_{reg} = \frac{\sum_{i=1}^{N}(1-\hat{x_i})}{N-1}; \quad \hat{x_i} = \frac{x_i}{max(x_i)} \tag{3.1}$$

where N is the number of tissues and x_i is the posterior probability profile component. Using the thresholds suggested in Yanai et al. (2005), the TSREs were categorised into low ($\tau_{reg} \leq 0.15$), intermediate ($0.15 < \tau_{reg} < 0.85$), high ($0.85 \leq \tau_{reg} < 1$) and absolute tissue-specific ($\tau_{reg} = 1$). This method was similarly applied to weak enhancer (state 6 posterior probability > 0.95) and weak promoter (state 1 posterior probability > 0.95) matrices.

3.3.5 Correlating TSREs with DHSs

DNase-Seq data from the ENCODE project was used to inspect the DNA accessibility of tissue-specific enhancers and promoters across multiple tissues. The mean of DNaseI signal was computed wherever multiple replicates were available within the ENCODE. The genomic coordinates of tissue-specific enhancers and promoters were compared with DNaseI hypersensitive hotspots using BEDTools (Quinlan, 2014) and the DNaseI signal in each tissue or cell line was extracted. I restricted the extraction of DNaseI signal to cases where 100% of the enhancer or promoter region overlapped the DHS hotspot, otherwise no DNaseI activity was assumed and a value of '0' was assigned to that enhancer or promoter. This resulted in a matrix of DNaseI signal corresponding to the posterior probability matrix of tissue-specific enhancers and promoters. To quantify the concordance between TSREs (tissue-specific enhancers and promoters) and DHSs, Pearson's correlation between posterior probability of their respective chromatin state and the corresponding DNaseI signal was calculated. The pairwise correlations between the tissues were visualised in a heatmap, and rows and columns were ordered based on hierarchical clustering (Fig. 3.9C-D).

3.3.6 Identifying SEs in the mouse genome

To identify SEs in the mouse genome, I implemented an approach similar to that previously used by Whyte et al. (2013). Using the ROSE algorithm, tissue-specific enhancers within a distance of 12.5 kb were stitched together into cohesive units and

ranked based on their H3K27ac signal. A TSS exclusion size of 2,000 bp was used to exclude tissue-specific enhancers within ±2 kb of known TSSs to remove any promoter bias. The algorithm calculates a threshold of the inflection point for H3K27ac signal (Fig. 3.10B). The stitched cohesive units with H3K27ac signal higher than the estimated threshold are defined as SEs while the remaining cohesive units are termed as TEs. The metagene profiles of mean H3K27ac signal across all the SEs and TEs (Fig. 3.10C) were generated using ngs.plot (Shen et al., 2014). Metagene plots are centered on the enhancers and display average ChIP-seq read density over all the enhancer regions and 2 kb window surrounding them. For visual comparison between profiles of SEs and TEs, the range of the y-axis were synchronised. For comparing the H3K4me1, H3K27ac and DNAseI hypersensitive signal over the stitched enhancers (Fig. 3.10D), the read density over these regions was calculated in reads per million. For H3K4me1 and H3K27ac ChIP-seq signal, the input control density was subtracted from the calculated read density. The read density for each feature was then normalised by dividing the signal at each enhancer by the maximum signal in each feature.

The non-parametric effect size (ES) was calculated as the difference in medians of the two groups divided by the pooled median absolute deviation (MAD). The following formula was used:

$$ES = \frac{Median_1 - Median_2}{MAD_{pooled}};$$

$$MAD_{pooled} = \sqrt{\frac{MAD_1^2 + MAD_2^2}{2}}; \tag{3.2}$$

$$MAD = median(|x - median(x)|)$$

3.3.7 Sequence conservation of mouse enhancers

The sequence conservation of SE and TE constituents was analysed using the phastCons scores for the mouse genome (mm9) across 20 mammalian species. The phastCons scores for the placental mammal subset (mouse, rat, guinea pig, rabbit, human, chimp, orangutan, rhesus, marmoset, bushbaby, tree shrew, shrew, hedgehog, dog, cat, horse, cow, armadillo, elephant and tenrec) was downloaded from the UCSC genome browser. The mean phastCons score for each constituent enhancer was calculated as: sum of the phastCons score of each bp in enhancer divided by the total length of the enhancer (i.e. 200). To generate a background set of regions (control), the location of enhancers were shuffled to random non-overlapping genomic positions such that the size, number and chromosome information of enhancers was preserved. The mean phastCons score

for the control set was calculated in a similar way to enhancers. This was repeated 1,000 times to produce a distribution of phastCons scores for the control regions. The p-values were computed using the Wilcoxon Rank Sum Test and the non-parametric effect size was calculated as described in section 3.3.6.

3.3.8 Enrichment of DA-SNPs in the mouse enhancers

The curated list of non-coding DA-SNPs reported in Hnisz et al. (2013) was used in this analysis. Hnisz et al. (2013) retrieved all the DA-SNPs in GWASs from the NHGRI database (Welter et al., 2014) and further selected the SNPs with: a dbSNP identifier; and those associated with a trait in at least two independent studies. For the comparison of these DA-SNPs with mouse enhancers, GWAS traits with at least 25 SNPs and those relevant to the mouse tissues in our dataset were selected, resulting in 26 phenotypic traits and diseases comprising of 1,592 SNPs. The coordinates of human SNPs (hg19) were converted to mouse genome (mm9) using UCSC liftover which resulted in 820 mouse homologous SNPs. For calculating the density of DA-SNPs in the mouse SEs and TEs, the number of DA-SNPs overlapping the SE and TE constituents in each tissue were counted. The count was then divided by the total genomic coverage (in bp) of SE and TE constituents in the corresponding tissue, and multiplied by a million to get the density in SNP per Mb. For a relative comparison, the density of DA-SNPs was also calculated in the human SEs and TEs. The coordinates of SEs and TEs in the human genome were retrieved from Hnisz et al. (2013) in 15 tissues and cell-types which were similar to the tissues in the mouse dataset for an unbiased comparison. These tissues and cell-types were: Adipose, B-cell, Brain Hippocampus Middle (Brain_HM), Brain Mid Frontal Lobe (Brain_MFL), Brain Inferior Temporal Lobe (Brain_ITL), GM12878, H1, Heart Left Ventricle (Heart_LV), Heart Right Ventricle (Heart_RV), Heart Right Atrium (Heart_RA), K562, Lung, SmallIntestine, Spleen, Thymus.

To calculate the significance of DA-SNPs overlapping the SE and TE constituents in the mouse genome, a permutation test was performed by shuffling the positions of DA-SNPs to random non-overlapping locations on the same chromosome. The density of the random SNPs in mouse enhancers was computed in a similar fashion to the real DA-SNPs. This process was repeated 1,000 times to produce a distribution of SNP density scores for background SNPs. The p-value was calculated by dividing the number of times the permuted SNP density was greater than the observed SNP density, by the total number of permutations. For all the DA-SNPs occurring in mouse enhancer regions, the permutation test shows that the overlap is significant ($p \leq 0.007$).

3.4 Discussion

In this chapter, I mapped potential regulatory elements in 22 mouse tissues by modelling multiple histone marks. Taking advantage of these annotations in a diverse range of tissues, I sought to identify tissue-specific regions to get better insights into tissue-type specific regulation. Previous studies have mostly utilised clustering techniques to identify shared and tissue-specific groups of regulatory elements across multiple tissues (Ernst et al., 2011; Roadmap Epigenomics Consortium et al., 2015; Shen et al., 2012). As an alternative to clustering, I implemented the Tau metric to calculate the tissue-specificity index of each regulatory element and systematically identified highly tissue-specific enhancers and promoters across 22 tissues. I believe this approach is a powerful method to identify and better quantify tissue-specific regions in a dataset. To gain a more profound understanding of enhancer regulation, I further classified tissue-specific enhancers into SEs (24%) and TEs (76%), henceforth generating a catalogue of different enhancer classes in 22 tissues, which includes previously unexplored tissues in the mouse.

Both SE and TE constituents appear to have low evolutionary conservation across the 20 mammalian species analysed here. However, SE constituents substantially lack sequence conservation compared to TEs, indicating SEs may have recently evolved possibly as a result of rapid functional evolution. But, it is difficult to confirm this solely on the basis of sequence conservation and further comparative analysis of functional data (such as that from ChIP-seq) across these mammalian species would be required. If true, these results would be consistent with the concept that key genes that regulate cell state have evolved to be regulated by SEs (Hnisz et al., 2015). To the best of my knowledge, only one study (Khan and Zhang, 2017) has analysed the sequence conservation of SE and TE constituents in two tissues, namely mESCs and pro-B cells. The authors compared the average phastCons score between SE and TE constituents (but not to the background regions) and identified SE constituents to have higher sequence conservation in pro-B cells. My dataset expands this analysis to a diverse range of tissues and cell lines, and also identifies a novel pattern which could not have been captured in a dataset comprising of a small number of tissues.

As more than 60% of the non-coding DA-SNPs in the human genome have been estimated to occur within enhancer regions (Hnisz et al., 2013), I investigated whether homologous positions of these SNPs in the mouse also overlap enhancers, which could help identify DA-SNPs with a greater potential to phenocopy in mouse models. Surprisingly, out of 1,592 non-coding DA-SNPs from GWASs, only $\sim$51% (820) were identified to have a homologous position in the mouse. Out of the 26 GWAS traits and diseases, DA-SNPs from16 traits were enriched in enhancers of human disease-relevant

tissues. DA-SNPs from 9 of these traits were also enriched in enhancers of mouse disease-relevant tissues. Whereas DA-SNPs from the remaining 7 traits occurred in enhancer regions specific to humans. Although *cis*-regulatory regions have diverged between the human and mouse genomes (Yue et al., 2014), some disease-associated non-coding regions appear to have conserved regulatory connections. Such regions could be potentially targeted to generate mouse models of DA-SNPs from GWASs. Taking into account such conserved regulatory pathways along with other epigenetic differences will help us to better translate the biomedical understanding acquired from mouse models into human research.

This study has a number of limitations that could influence my results. First, the association between histone modifications and enhancer activity is not completely comprehended as even combinations of histone marks do not correlate perfectly with enhancer occupancy (Arnold et al., 2013; Bonn et al., 2012). Moreover, in the scientific community, there has been a lack of consensus in histone marks being used for enhancer prediction (Shlyueva et al., 2014). Recent studies have used only H3K27ac as a mark of active enhancers, as opposed to the combination of H3K27ac, H3K4me1 and H3K4me3 used in earlier studies. The enhancer prediction strategies appear to be dependent on the availability of the data instead of standard conceptual models. Furthermore, a recent study (Pradeepa et al., 2016) demonstrated novel acetylation of lysine residues (H3K64ac and H3K122ac) to correlate with active promoters and enhancers. Therefore, it is still unknown what combination of histone marks would be sufficient to capture a high degree of enhancer activity. In this study, only 72% of the mapped strong enhancers overlapped active DHSs and the known TFBS had a moderate enrichment within the strong enhancers, which indicates that using only histone marks to predict the location of enhancers may not be the most accurate method. An alternative strategy would use open chromatin regions captured from DNase-seq (Boyle et al., 2008) or ATAC-seq (Buenrostro et al., 2015), which have been proven to unbiasedly predict all regions with regulatory activity (Thurman et al., 2012). Since open chromatin regions directly overlay regions with regulatory activity, DNase-seq/ATAC-seq produces narrow peaks with their summit over the core TF occupancy within the regulatory region. This allows open chromatin data to more accurately predict the location of regulatory elements as opposed to histone modifications, which occur at the regions flanking the regulatory elements and hence produce very broad peaks. At the time of analysis, DNase-seq or ATAC-seq data was only available in a limited number of mouse tissues and hence was not implemented in the enhancer prediction strategy.

A second limitation of this study is that histone modification data cannot directly identify which TFs are bound within these regulatory elements. Again, DNaseI accessibility data could also help in overcoming this limitation. DNase-seq digital footprints

generated at high sequencing depths can detect TF binding at near base-pair resolution (Hesselberth et al., 2009; Neph et al., 2012). Such datasets have been modelled to computationally predict the effect of non-coding regulatory variants on TF binding (Schwessinger et al., 2017). A third limitation focuses on the dataset. Since the dataset analysed here comprises of only 22 tissues, some enhancers with high tissue-specificity may shift towards intermediate tissue-specificity if more tissues are included. Indeed, applying this method to more tissues would further refine the tissue-specific profiles of enhancers generated here. Nevertheless, this data provides a platform to conduct further analysis of transcriptional regulation and better understand the role of enhancers in disease aetiology.

Chapter 4

Impact of enhancer architecture on gene function and mouse phenotypes

In this chapter, I investigate SE and TE properties, how they influence gene expression and their involvement in disease aetiology. A section of this chapter describes the enrichment of TF ChIP-seq peaks within enhancers (section 4.2.5), which was carried out in collaboration with the Makeev lab at the Vavilov Institute of General Genetics, Moscow. The results described in this chapter contributes towards the following article:

Sethi, S., I. E. Vorontsov, I. V. Kulakovskiy, S. Greenaway, J. Williams, V. J. Makeev, S. D. M. Brown, M. M. Simon, A.-M. Mallon (2019).“Deciphering the impact of enhancer architecture on gene function and mouse phenotypes”. Under review in *Cell Reports*.

4.1 Introduction

The identification of enhancer regions in the genome is just the first step towards studying the transcriptional regulation. Investigating these potential enhancer regions to understand gene regulatory networks and mechanisms still remains a challenge. As discussed in chapter three, more than 60% of the disease-associated genetic variation from GWASs occurs within potential enhancer regions. Disruptions to any of the enhancer regions may lead to disease in humans and related phenotypes in model organisms such as the mouse (Bhatia and Kleinjan, 2014; Kleinjan and Lettice, 2008; Maston et al., 2006). The number and scale of these observations from GWASs has driven research to characterise enhancers and their association to pathological states.

Until recently, studying the effect of enhancer disruption *in vivo* has been difficult. Though reporter assays in cultured cells have been useful to show that a stretch of sequence can function as an enhancer outside of their native environment (Patwardhan et

al., 2012), they cannot show if the enhancer is responsible for initiating the transcription of a particular target gene *in vivo*. However, studying enhancers and other non-coding elements *in vivo* has been greatly facilitated by CRISPR-Cas9 and on a case-by-case basis we are beginning to understand the roles of enhancers in the susceptibility and pathogenicity of diseases (Canver et al., 2015; Cunningham et al., 2018; Diao et al., 2016; Dickel et al., 2018a; Groschel et al., 2014; Korkmaz et al., 2016; Li et al., 2014; Moorthy and Mitchell, 2016; Rajagopal et al., 2016; Seruggia et al., 2015). The most common approach has been to genetically manipulate the enhancer either by deleting or disrupting it, and examine its effect on gene expression *in vivo*. Such studies have demonstrated that deletion of critical enhancer sequences can abolish almost completely (Canver et al., 2015) or up to 90% (Groschel et al., 2014; Li et al., 2014) of their target gene expression, whereas some enhancers could be partially redundant and cause subtle changes to the expression (Cunningham et al., 2018; Moorthy et al., 2017; Osterwalder et al., 2018). Even single base pair changes within enhancers have been shown to drastically effect the target gene expression, leading to disease susceptibility (Bauer et al., 2013; Mansour et al., 2014). However, the degree to which enhancers contribute to target gene expression on a global scale remains indistinct, as they have been mostly studied at a few individual genomic loci.

Another question which arises about enhancer function is whether a small number of enhancers are enough to activate transcription or whether enhancers share their role across a large number of enhancer elements. The concept of SEs proposes that dense clusters of enhancers represent a new paradigm in gene regulation, having a greater role in the control of mammalian cell state. Systematic mapping of SEs using H3K27ac chromatin mark across diverse human tissues and cell lines has shown that SEs regulate key genes that define the cell identity, including known master regulators (Hnisz et al., 2013; Huang et al., 2016; Loven et al., 2013; Pelish et al., 2015). For instance, SEs in mESCs were detected to regulate key pluripotency genes (such as *Oct4*, *Sox2* and *Nanog*) and have higher TF binding for *Klf14* and *Esrrb* (Hnisz et al., 2013). Furthermore, SEs have been identified to drive high total-expression (aggregated expression of all exons) of their target genes compared to TEs in a wide range of human cell-types (Hnisz et al., 2013). While studies in the mouse genome find similar results, they are currently less comprehensive and limited to relatively few tissue types (Adam et al., 2015; Fang et al., 2015; Ohba et al., 2015; Shin et al., 2016; Siersbæk et al., 2014; Vahedi et al., 2015; Whyte et al., 2013). In addition to this total-expression, a few studies have shown SEs to be associated with tissue-specific gene expression (tendency of a gene to be specifically expressed in a tissue or cell line) in cell lines. For instance, genes associated with SEs in multiple myeloma cell line had a tendency to be specifically expressed in myeloma cells (Loven et al., 2013). Moreover, as shown in the previous chapter, SEs in human cell-types frequently harbour disease-causing

variation (Farh et al., 2015; Hnisz et al., 2013; Parker et al., 2013), while TEs have been considered less important. However, to date there has been no systematic study defining genome-wide functional differences between SEs and TEs, and their relationship to phenotypes.

Although, SEs have been characterised in multiple cell-types, it is debatable whether the concept of SE represents a novel model of regulatory class or simply a cluster of traditional enhancers. An important question is whether the activity of individual enhancers depends or is affected by other enhancers in the cluster or in the nearby vicinity. It remains unclear whether the individual components in a SE, referred to as constituent enhancers, work in an additive manner (independent activity) or have a more complex cooperative effect (dependent activity) on gene expression of their target genes. Several recent studies investigating the function of constituent enhancers at various individual genomic loci have shown contrasting results (Hay et al., 2016; Hnisz et al., 2015; Shin et al., 2016; Suzuki et al., 2017). *In vivo* studies at *Wap* and α-globin associated SE locus show that deleting individual constituent enhancers have variable effect on target gene expression and most of the single constituent enhancer deletions do not terminate the enhancer function (Hay et al., 2016; Shin et al., 2016). However, combinatorial deletions of constituent enhancers result in a higher reduction of target gene expression, hence demonstrating that constituent enhancers work independently and exhibit more or less an additive effect (Hay et al., 2016; Shin et al., 2016). Whereas, reporter assay tests at the *Pou5f1* SE locus show that combinatorial deletions of constituent enhancers often cause a lower reduction in target gene expression compared to deletion of a single constituent enhancer, hence demonstrating a complex influence of constituent enhancers on each others activity (Hnisz et al., 2015). Furthermore, a recent study which performed deletions of constituent enhancers in multiple cell lines showed that the majority of the constituent enhancer deletions can cause drastic reduction (50%-80%) in target gene expression (Suzuki et al., 2017). Therefore, further studies are required to understand whether these observations are exceptions or represent a genome-wide pattern.

In this chapter, I aim to study the relationship and functional impact of enhancer architecture on gene function and mouse phenotypes (Fig. 4.1). Using the genome-wide enhancer maps produced in chapter three, and gene expression profiles in 22 mouse tissues, I systematically analyse the influence of different enhancer types on total-expression and tissue-specificity of their target genes. Additionally, I model the relationship between constituent enhancer density and target gene expression on a global scale. I further explore and compare the mammalian phenotypes and disease traits associated with potential target genes associated with different enhancer types. Using standardised mouse phenotyping data from gene knockout models, I analyse the phenotype severity and pleiotropic effects of SE and TE associated gene knockouts.

Finally, I go on to model regulatory data along with other molecular characteristics to infer mammalian gene-phenotype associations and to identify potential novel pathogenic genes which may be used for further characterisation. Overall, the results from this chapter show that genes harbouring different enhancer architecture tend to have distinct expression patterns, but contribute to phenotype outcomes at a comparable level, thus providing novel insights to enhancer-phenotype relationship.

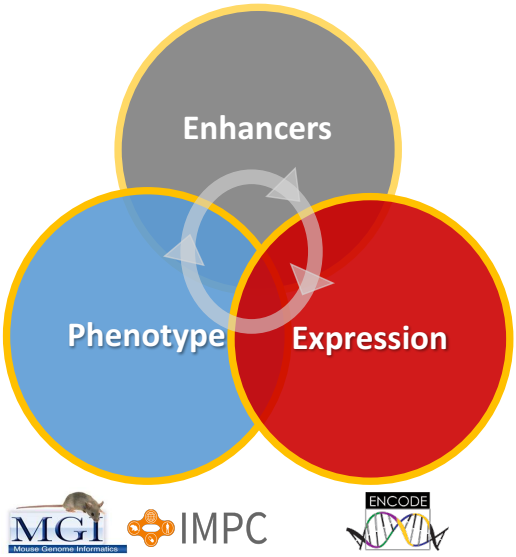


Fig. 4.1 Research aims. A schematic displaying the relationships I aim to investigate in this chapter. The gene expression data in various tissues was collected from the ENCODE project and mouse phenotypes associated with the genes were retrieved from the Mouse Genome Database (MGD) (Blake et al., 2017) and the IMPC.

4.2 Results

4.2.1 Associating regulatory elements to potential target genes

Enhancers have been identified to interact with promoters of adjacent genes through looping in order to activate their transcription (Gondor and Ohlsson, 2009; Ong and Corces, 2011; Sanyal et al., 2012; Spitz and Furlong, 2012). Previous investigations have likewise observed SEs to frequently overlap the genes they regulate (Hnisz et al., 2013; Whyte et al., 2013). A previous study in murine ESCs identified more than 80% of SEs and TEs to interact with their nearest active gene using ChIA-PET (Dowen et al., 2014). In order to explore the functional role of TSREs, I associated each element to a potential target gene using GREAT (McLean et al., 2010). GREAT defines computationally derived 'regulatory domains' in the genome and associates non-coding regions to their likely target genes in the same domain. Overall, this approach identified 3,617 and 14,832 protein-coding genes associated with SEs and TEs in at least one tissue or cell-type, respectively. The resulting enhancer-gene associations were highly consistent with previously identified TADs in the mouse genome (Dixon et al., 2012); 96% of the enhancer-gene pairs (where both the enhancer and its target gene overlapped a TAD) were identified to be in same cortex TADs and 93% in the same mESC TADs (Fig. 4.2A). Similarly, 87% of the enhancer-gene pairs were identified in the same computationally derived enhancer-promoter units (EPUs) in the mouse genome (Shen et al., 2012). As expected, the majority of the tissue-specific enhancers (62.53% of SEs, 57.25% of TEs) are located within 50 kb from the TSSs of their associated genes, while 42% of active promoters are associated with TSSs within 5 kb (Fig. 4.2B). These observations are in agreement with previous findings by Chepelev et al. (2012), where 55.80% (1,324/2,373) of the enhancer-promoter interactions detected using ChIA-PET were identified to occur within < 50 kb of the promoters. These predicted SEs, TEs and their potential target genes were used for all the subsequent investigations.

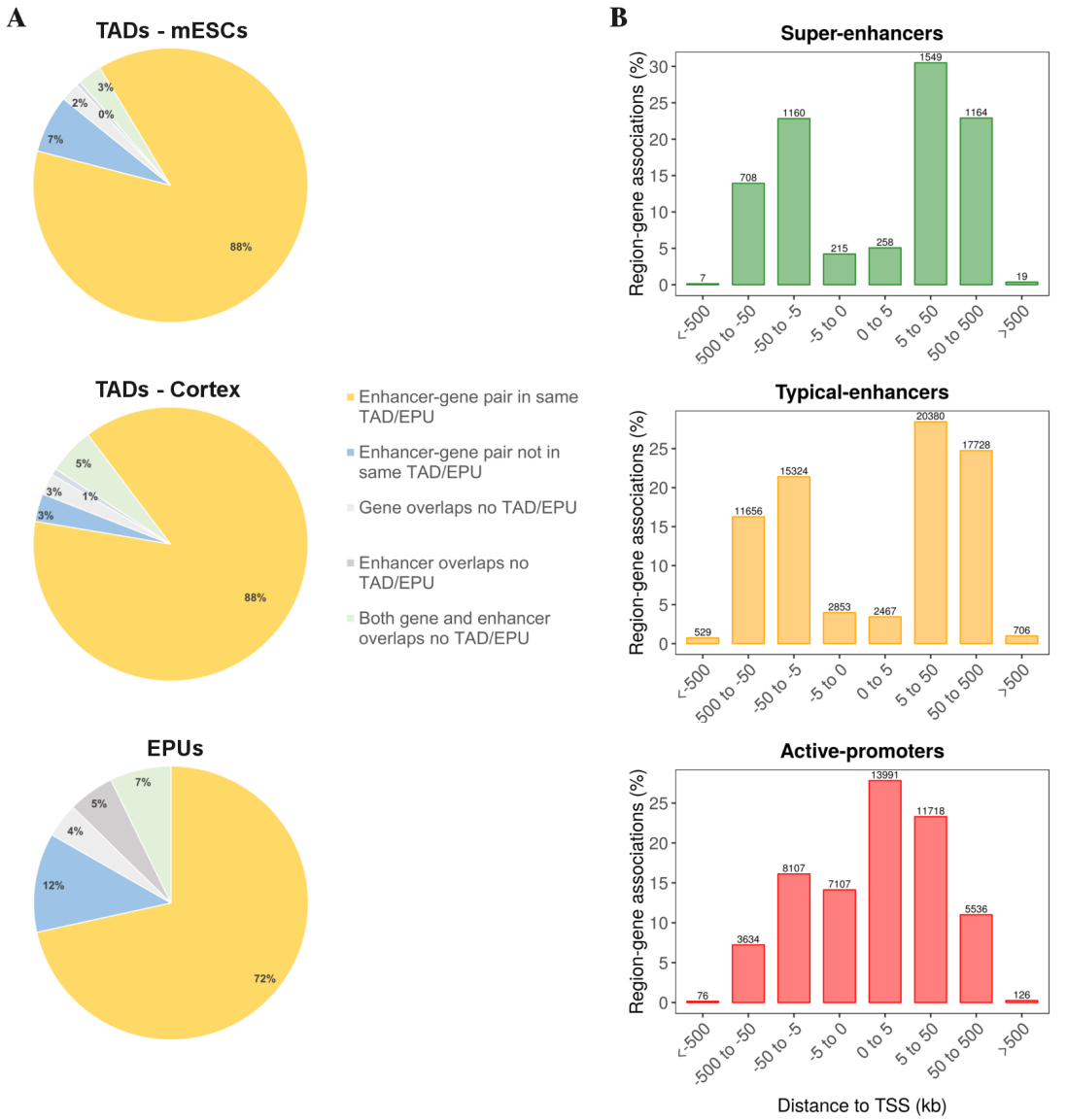


Fig. 4.2 Region-gene associations of regulatory elements. (A) Pie charts displaying the proportion of gene-enhancer pairs within previously reported TADs and EPUs in the mouse genome. (B) Bar plots binned by orientation and TSS showing the distance between various regulatory elements and their putative target genes. For all three graphs, the y-axis represents the percentage of region-gene associations while the number of associations in each bin are listed in the graph. The x-axis shows the distance (divided into separate bins) of the region relative to the TSS of the gene. Negative distance depicts regions upstream of the TSSs; positive distance depicts regions downstream of the TSSs; 0 represents the TSSs.

4.2.2 Profiling genome-wide enhancer activity and target gene expression

Influence of enhancer type on target gene expression

Prior investigations have shown SEs to be related with highly expressed genes in multiple human cell-types (Hnisz et al., 2013). In addition, a few studies have shown SEs to be associated with tissue-specific gene expression (Loven et al., 2013). With the aim of exploring whether this association prevails across multiple tissue types and different enhancers, I examined the impact of these newly identified enhancers in 22 tissues. In order to inspect the relationship between various enhancer types and the expression of genes they potentially regulate, I utilised ENCODE's RNA-seq data. To effectively identify any common expression patterns between genes, tissues and enhancers, I constructed a dataset formed of genes expressed within a particular tissue, termed gene-tissue pairs, followed by categorisation on their type of enhancer association, hence grouping them into three classes: (1) gene-tissue pairs associated with SEs, referred to as super-enhancer class (SEC); (2) gene-tissue pairs associated with TEs, referred to as typical-enhancer class (TEC); and (3) gene-tissue pairs associated with weak/poised enhancers, referred to as weak-enhancer class (WEC).

I found that both the SEC and TEC are associated with highly expressed genes in comparison to the WEC (SEC: ES = 0.95, $p < 2.2 \times 10^{-16}$; TEC: ES = 0.86, $p < 2.2 \times 10^{-16}$; Wilcoxon Rank Sum Test) but that the SEC appears to have the highest level of total-expression (SEC compared to TEC: ES = 0.56, $p < 2.2 \times 10^{-16}$) (Fig. 4.3A, Appendix A.7A). Likewise, I investigated whether these newly identified enhancers in mouse tissues drive tissue-specific expression of their associated genes. The SEC is observed to have higher tissue-specific expression (quantified as $\tau_{exp-frac}$, see methods 4.3.3) compared to the TEC (ES = 0.62, $p < 2.2 \times 10^{-16}$; Wilcoxon Rank Sum Test) and WEC (ES = 0.96, $p < 2.2 \times 10^{-16}$) (Fig. 4.3B). However, not all genes in the SEC display tissue-specific expression. To further understand tissue-specific expression of the genes within different enhancer classes, I categorised it into three levels of low, intermediate and high (see methods 4.3.3). This identified 16.46% (690/4,191) of the SEC, 4.42% (1,923/43,484) of the TEC and 3.38% (230/6,795) of the WEC to have high tissue-specific expression ($\tau_{exp-frac} \geq 0.85$) (Fig. 4.3C, Appendix A.7B). However, further examination of the high tissue-specific expression category shows that the absolute number of genes within the TEC (1,923) is notably higher than the SEC (690) and the WEC (230). Overall this data shows that the ratio of genes within the SEC with high tissue-specific expression is at least 4 times larger than the genes within other enhancer classes. However, its absolute number is smaller compared to the

TEC, which contributes the largest amount (68%) of enhancer associated tissue-specific expression in the genome (Fig. 4.3D) and should also not be overlooked.

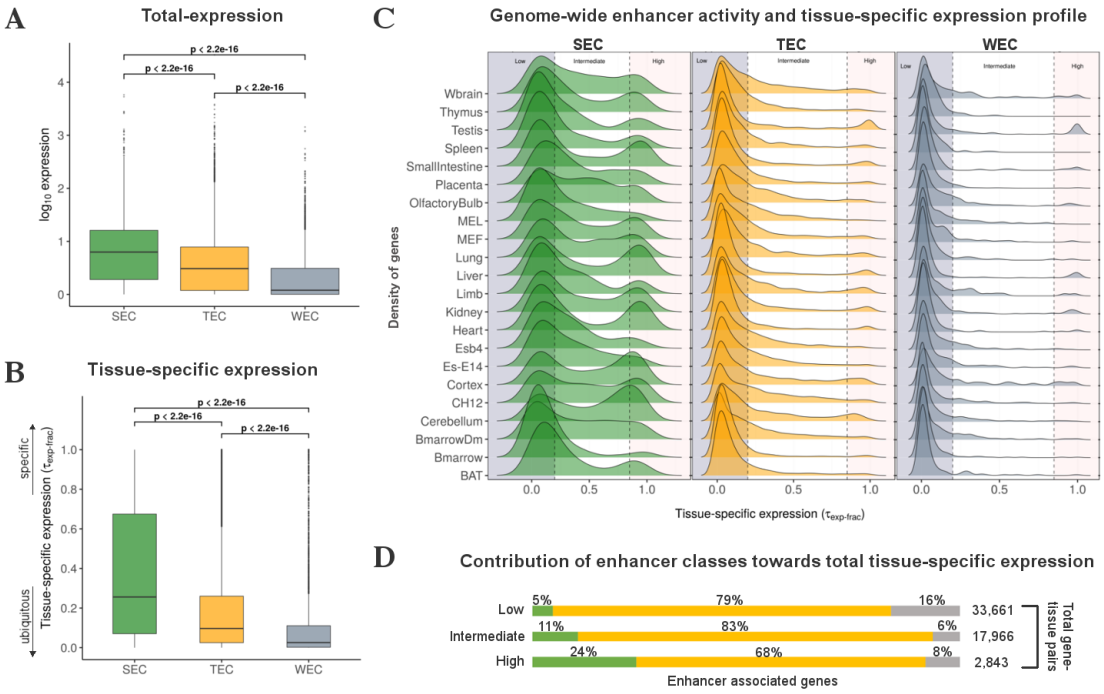


Fig. 4.3 Enhancer activity and its influence on gene expression. (A) Box plot showing the total-expression (in log-transformed RPKM) of different enhancer classes across 22 tissues. Each box plot shows: the median, middle bar; interquartile range, the box; 1.5 times the interquartile range, the whiskers. (B) Box plot showing the tissue-specific expression (as measured by $\tau_{exp-frac}$) of different enhancer classes across 22 tissues. The p-values were calculated using the Wilcoxon Rank Sum Test. (C) Density plots showing the distribution of genes within tissue-specific expression categories (low, intermediate, high) in different enhancer classes. The Y-axis for each tissue displays the density of genes scaled across the tissues but not across the enhancer classes. (D) Contribution of each enhancer class (in percentage) towards the total number of enhancer associated genes in the genome, categorised by their tissue-specific expression. See also Appendix A.7.

Impact of constituent enhancer density on target gene expression

SEs are likely to be comprised of large number of constituent enhancers (Fig. 4.4A). The average number of constituents enhancers within SEs is 13 compared to only 3 in TEs. Therefore, I asked whether constituent enhancer composition have any impact on total- and/or tissue-specific expression of their associated genes. To investigate this, I combined both SEs and TEs into a single dataset. I compared the frequency of the constituent enhancers (total number of constituent enhancers associated with a gene) within the combined dataset with total- and tissue-specific expression of their associated genes, which revealed a significant, but weak positive correlation respectively (total-expression: Spearman correlation $r = 0.12$, $p < 2.2 \times 10^{-16}$; tissue-specific expression: $r = 0.18$, $p < 2.2 \times 10^{-16}$) (Fig. 4.4B-C). In contrast, weak-enhancers show little to no correlation with total-expression ($r = -0.03$, $p = 0.02$) or tissue-specific expression $r = 0.04$, $p = 0.001$) of their associated genes (Fig. 4.4D-E). Furthermore, using the tissue-specific expression categories (low, intermediate and high), 31% of genes with high tissue-specific expression were identified to be associated with 10 or more constituent enhancers as opposed to 15% in genes with intermediate tissue-specificity and 8% in genes with low tissue-specificity, showing that genes with high tissue-specific expression tend to be related with relatively greater number of constituent enhancers (Fig. 4.4F). Overall this shows that total- and tissue-specific expression modestly increases with the number of constituent enhancers, which could indicate a non-additive relationship between them. A weak correlation could suggest the existence of cooperative activity or partial redundancy amongst the constituent enhancers. Hence, this analysis predicts that constituent enhancers may exert a complex, instead of a simple additive effect on the transcriptional output. However, it should be noted that this is a computational prediction and has limitations. In order to accurately calculate the impact of constituent enhancers on target gene expression, it is important to know which constituent enhancers are real/active and which gene(s) they precisely regulate. The presence of false-positives in these variables can have a significant effect on the prediction from this analysis.

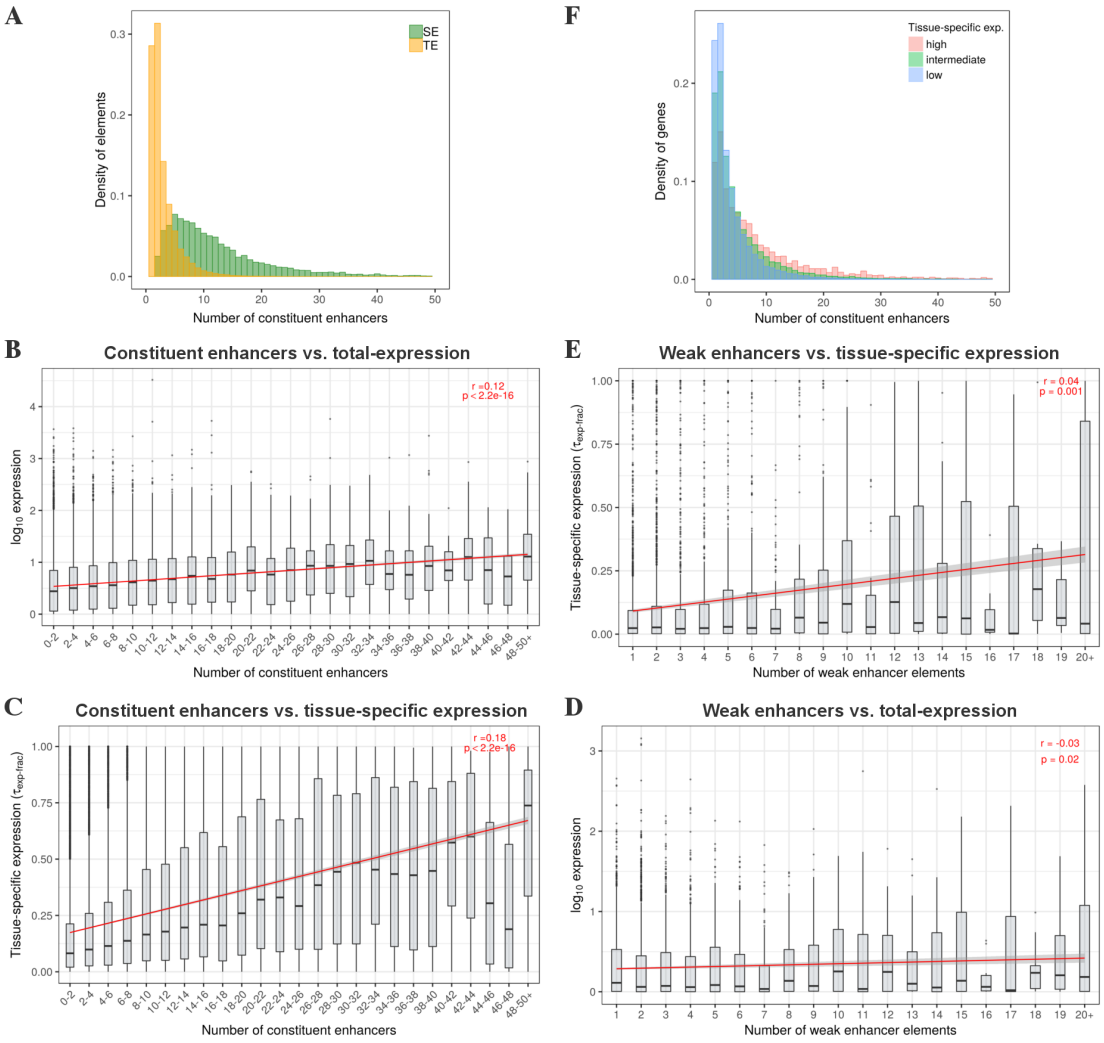


Fig. 4.4 Impact of constituent enhancer density on target gene expression. (A) Distribution of constituent enhancers within SEs and TEs across all 22 tissues. (B-C) Correlation analysis between number of constituent enhancers (within SEs and TEs) and (B) total-expression; (C) tissue-specific expression of their associated genes. (D-E) Correlation analysis between number of weak enhancers and (D) total-expression; (E) tissue-specific expression of their associated genes. Correlation coefficient (r) was calculated using Spearman rank correlation. (F) Distribution of enhancer associated genes as a function of number of constituent enhancers within their associated SEs and TEs. The genes are categorised based on their tissue-specific expression into: highly tissue-specific, intermediate tissue-specific and low tissue-specific.

Impact of multiple enhancer-gene associations on tissue-specific expression

Since a gene could be related to active SEs or TEs in multiple tissues, not necessarily at the same genomic location across the different tissues, I inspected these multiple enhancer-gene associations and their effect on tissue-specific expression. For this purpose, I assessed the number of distinct tissues, where an enhancer associated with a gene occurs, which I define here as 'enhancer tissue-types' (Fig. 4.5A). A large portion ($\sim$78%, 2,838 out of 3,617) of the SEC is associated with one enhancer tissue-type, i.e. the genes are associated with SEs of only one tissue. However, only $\sim$27% (3,955 out of 14,832) of the TEC have one enhancer tissue-type, while the remaining 73% of genes in the TEC are associated with TEs of two or more tissues (Fig. 4.5B). Furthermore, the genes with a high number of enhancer tissue-types are observed to be associated with low values of $\tau_{exp-frac}$ (Fig. 4.5C), hence increasing enhancer tissue-type association increases ubiquitous expression.

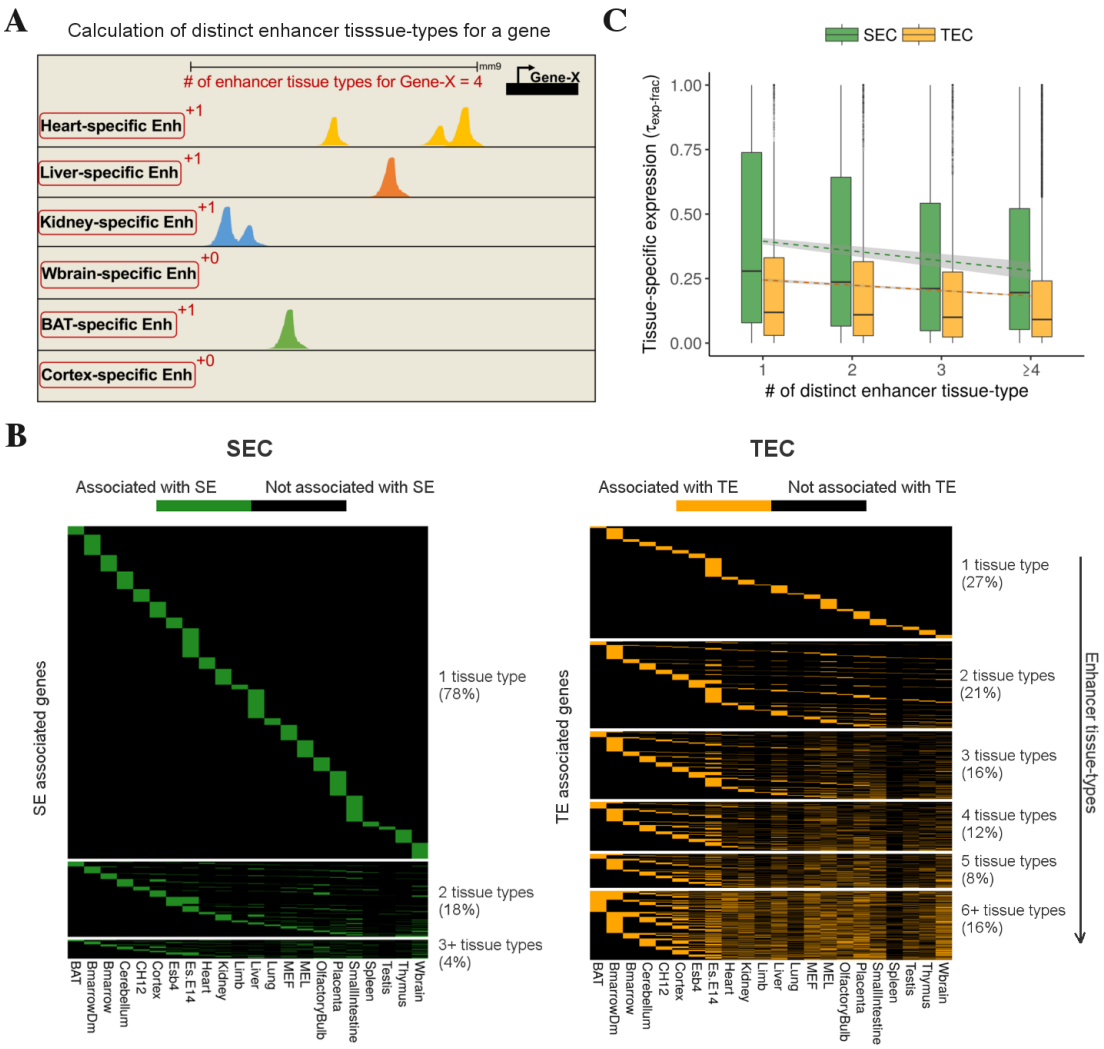


Fig. 4.5 Distinct enhancer tissue-types associated with genes. (A) A schematic to illustrate the calculation of enhancer tissue-types for a gene. The number of distinct tissue types of various enhancers associated with the gene of interest are added to compute the number of enhancer tissue-types. (B) Heatmaps displaying the enhancer tissue-types of SEC and TEC. Each row is a gene and columns represent its association with enhancers across 22 tissues and cell-types. (C) Box plot showing the correlation between the number of enhancer tissue-types and tissue-specific expression of SEC and TEC. The trend lines (green: SEs; orange: TEs) were calculated using linear regression.

4.2.3 Influence of enhancer architecture on phenotypes

Previous studies have identified SEs to be associated with genes that regulate cell state and therefore unlikely to be involved in a housekeeping role (Hnisz et al., 2013; Whyte et al., 2013). To explore the functional role of SE and TE associated genes in my dataset, I performed Gene Ontology (GO) enrichment analysis in 22 mouse tissues. Genes associated with SEs belonging to the SEC category are enriched for transcription factor binding activity ($p = 10^{-10}$), regulation of cell differentiation ($p = 10^{-23}$) and regulation of cell development ($p = 10^{-16}$) (Appendix B.1). The breadth of this analysis demonstrates novel cell identity associations in unexplored tissues in the mouse. As expected, SEs are associated with genes known to be important in the control and regulation of tissue or cell identity. Some examples of these novel SE associated genes include *Ucp1* (responsible for generating body heat in mammals (Cannon and Nedergaard, 2004)), *Pparg* (involved in differentiation of brown adipocytes (Nedergaard et al., 2005)), and the key TF *Ebf2* (which determines the fate and function of brown fat cells (Rajakumari et al., 2013)) in BAT; key TFs like *Gata4*, *Nkx2-5* and *Myocd* (critical for heart development and regulation of cardiomyocytes (Akazawa and Komuro, 2003)) in heart; *Cxcr2* (which regulates the emigration of neutrophils from bone marrow (Martin et al., 2003)) in bone marrow; and *Rbfox3* (splicing regulator of neuronal transcripts (Kim et al., 2009; Kim et al., 2013)) in cerebellum. Previously well studied master TFs in mESCs such as *Pouf51* (also known as *Oct4*), *Sox2*, *Klf4*, *Esrrb*, and *Prdm14* were also identified to be associated with SEs. On the other hand, the TEC appear to have different enrichments in GO analysis and are linked with genes involved in nucleotide and protein containing-complex binding ($p = 10^{-6}$), cellular protein localisation ($p = 10^{-7}$) and cell morphogenesis ($p = 10^{-5}$) (Appendix B.2). The TEC also includes a weak enrichment of transcription co-activator activity ($p = 10^{-2}$), though no association with transcription factor binding activity is observed. Furthermore, the TEC is significantly enriched for housekeeping genes ($p = 2.7 \times 10^{-11}$, Odds Ratio (OR) = 1.49, 95% CI [1.32, 1.68]), while the SEC is depleted ($p = 0.012$, OR = 0.82, 95% CI [0.69, 0.98]).

Phenotype associations of enhancer associated genes

In order to explore the role of enhancers in phenotypes, I investigated the mammalian phenotypes and human diseases associated with the genes within the SEC and TEC. For this purpose, I used the ToppFun tool (Chen et al., 2009) to calculate the enrichment of mammalian phenotype terms and human disease annotations in the SEC and TEC of each tissue individually. For computing the enrichments, ToppFun uses mammalian phenotype terms from the Mouse Genome Database (MGD) (Blake et al., 2017) and hu-

man disease annotations from databases like ClinVar (Landrum et al., 2014), DisGeNet (Pinero et al., 2017), GWASs and OMIM (Hamosh et al., 2005). Significant enrichment of both mammalian phenotypes and disease annotations in the corresponding tissue types was identified (Fig. 4.6 and Appendix B.3), suggesting a strong relationship between both SEC and TEC, and the resulting pathological outcomes (disease causation). For instance, genes associated with cerebellum-specific enhancers are enriched for phenotypes

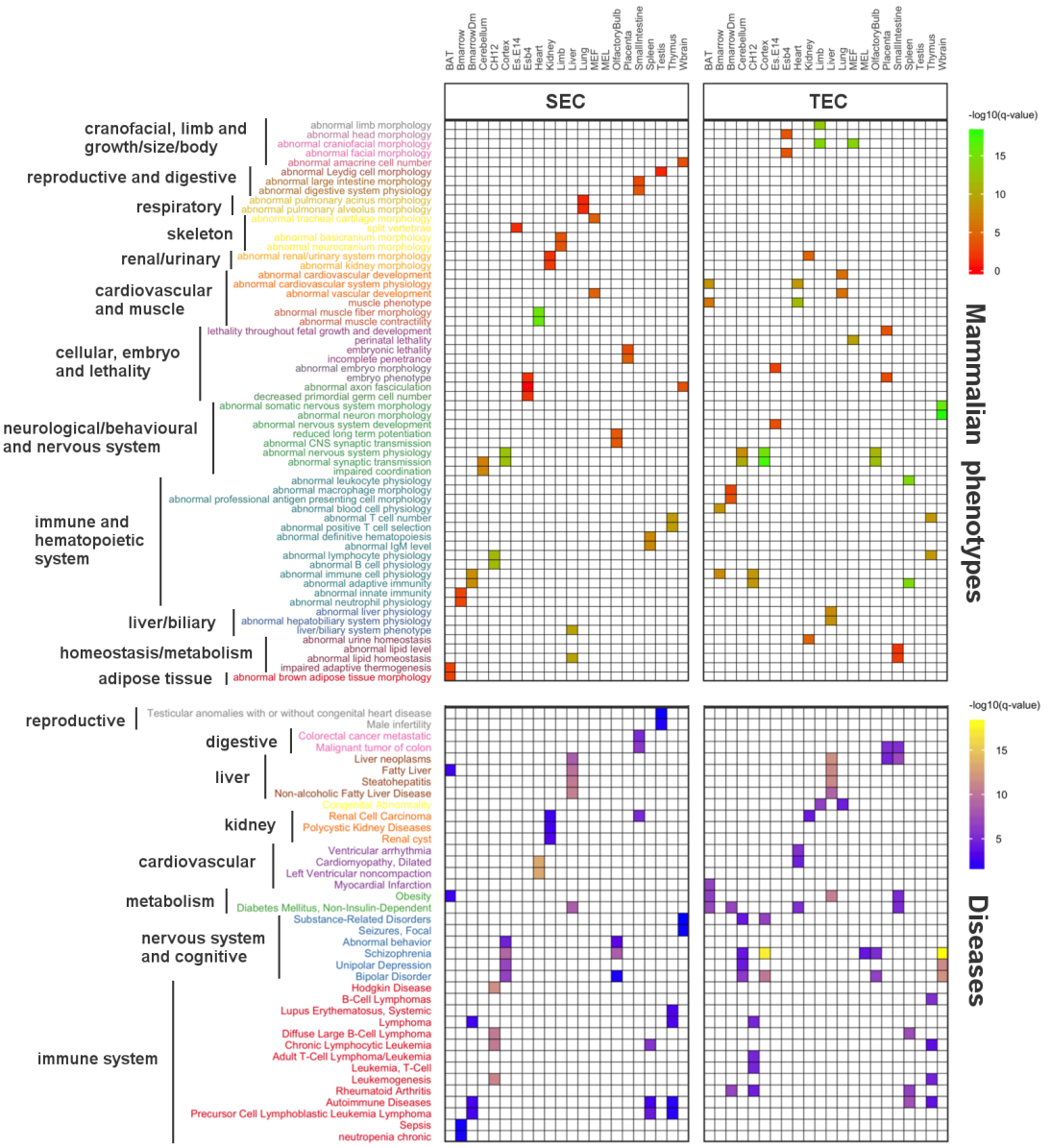


Fig. 4.6 Mammalian phenotype and human disease annotations enriched in the SEC and TEC. Heatmap displaying the most enriched mammalian phenotypes and human diseases amongst the SEC and TEC in each tissue. The cells in the heatmap display the FDR (q-value) associated with the enriched terms and was calculated using the Benjamini-Hochberg method. The enrichment analysis was performed using ToppFun, which retrieves mammalian phenotype annotations from the MGD and human disease annotations from ClinVar, DisGenNet, GWASs and OMIM. See also Appendix B.3.

such as impaired coordination ($q = 4.83 \times 10^{-8}$) and abnormal synaptic transmission ($q = 2.46 \times 10^{-7}$), and diseases such as bipolar disorder ($q = 8.52 \times 10^{-7}$) and unipolar disorder ($q = 6.26 \times 10^{-5}$). Likewise, genes related to heart-specific enhancers are enriched for phenotypes like abnormal cardiac muscle contractility ($q = 9.05 \times 10^{-16}$) and diseases like cardiomyopathy ($q = 5.45 \times 10^{-14}$). Furthermore, enrichment of blood-related cancers (such as Hodgkin Disease, $q = 1.90 \times 10^{-12}$; T-cell Leukemia, $q = 1.41 \times 10^{-5}$) in CH12 enhancer associated genes is consistent with the idea that oncogenes are placed under the effect of strong enhancers during cancer development leading to over-expression of these genes (Loven et al., 2013; Mansour et al., 2014). On the other hand, the WEC display either an insignificant or a weak association with phenotypes in the majority of the tissues (Appendix B.4). These findings confirm the role of tissue-specific enhancers in active gene regulation, tissue function and disease development.

As a validation to my above approach, i.e. examining the enrichment of mammalian phenotypes in enhancer associated genes, I likewise calculated its inverse, i.e. enrichment of enhancer-associated genes in mammalian phenotypes. For this analysis, the genes associated with various mammalian phenotypes were extracted from the MGD. The mammalian phenotypes were most notably enriched in the SEC and TEC of phenotype related tissue-types, subsequently reproducing the previously observed relationship amongst SEC and TEC with mammalian phenotypes (Fig. 4.7). For instance, genes associated with nervous system phenotype are most enriched for SEC and TEC in the whole brain, cortex and cerebellum. Next, I sought to examine the prevalence of enhancer associated genes in specific mouse phenotype traits. For this purpose, mouse phenotyping data was collected from the IMPC for several standardised phenotype procedures. Genes showing phenodeviancy for open field test ($n = 299$), grip strength ($n = 216$), body composition ($n = 206$), hematology ($n = 384$), acoustic startle and pre-pulse inhibition ($n = 151$), and heart weight ($n = 76$) were compared to enhancer associated genes to calculate their enrichment. These phenotype procedures and tests performed in the IMPC are described in the IMPReSS database (https://www.mousephenotype.org/impress). Contrary to the enrichment observed using the MGD data, no notable enrichment of enhancer associated genes is observed amongst these specific phenotype attributes from the IMPC (Fig. 4.8). A possible explanation for this observation could be the modest number of genes in the IMPC database at the moment, especially compared to the MGD which comprises of gene-phenotype associations from all genetic studies and not just gene knockouts.

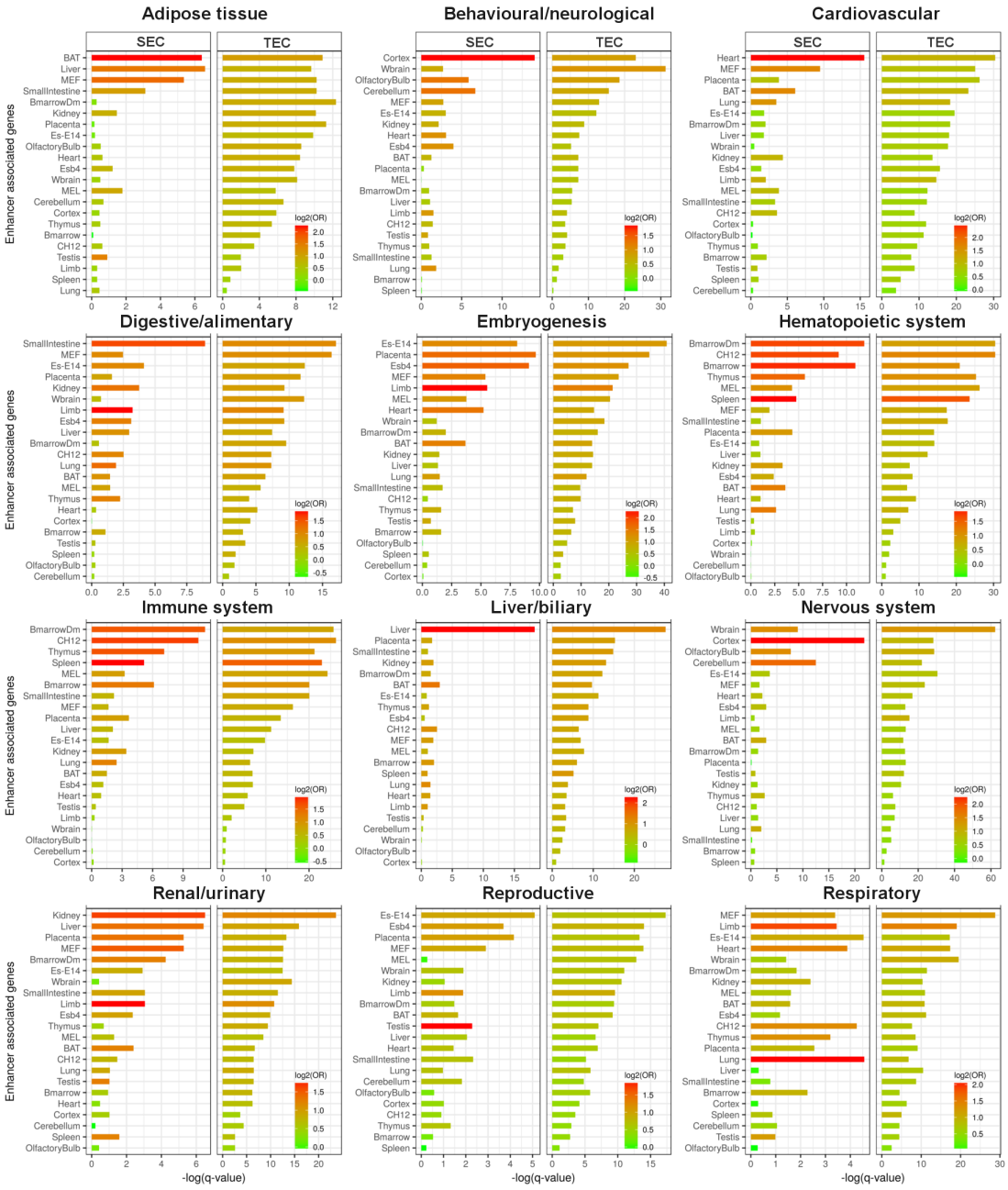


Fig. 4.7 Enrichment of enhancer-associated genes in mammalian phenotypes. Bar plots displaying the enrichment of SEC and TEC amongst genes associated with mammalian phenotypes in the MGD. The enrichment p-values and ORs were calculated using the Fisher's exact test. The p-values were further corrected for multiple testing (q-value) using the Benjamini-Hochberg method.

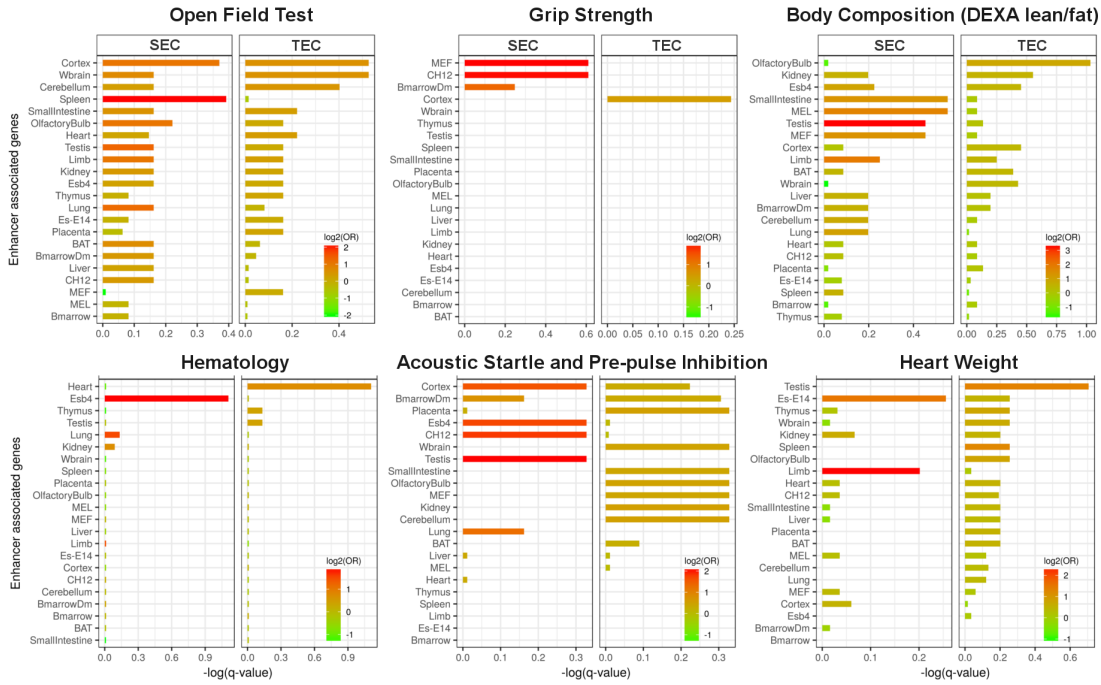
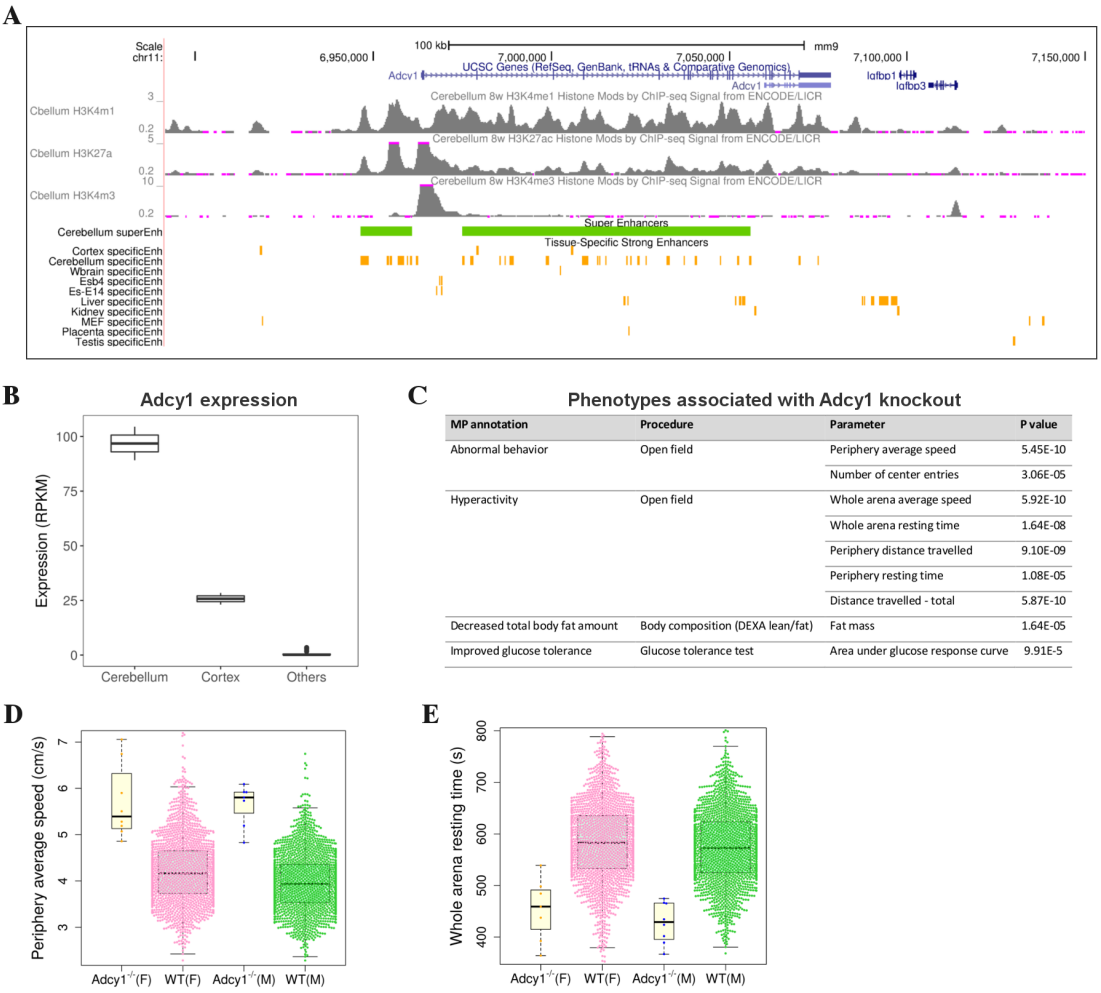


Fig. 4.8 Enrichment of enhancer-associated genes in IMPC phenotypic traits. Bar plots displaying the enrichment of SEC and TEC amongst genes associated with various phenotype procedures in the IMPC. The enrichment p-values and ORs were calculated using the Fisher's exact test. The p-values were further corrected for multiple testing (q-value) using the Benjamini-Hochberg method.

Examples of enhancer-phenotype correlation

Despite no significant enrichment of enhancer associated genes within the IMPC phenotypes, I inspected a few genes whose phenotype data from the IMPC correlates with the corresponding enhancer activity. In order to explain the enhancer-phenotype relationship, I describe the epigenomic landscape of two genes in the following section.

The first example describes the relationship between enhancers and mouse phenotypes for the gene Adenylate Cyclase 1 (*Adcy1*, MGI:99677). *Adcy1* is a neural-specific protein which converts ATP to cAMP (cyclic adenosine monophosphate) (Xia et al., 1993) and has been implicated in memory, learning, synaptic plasticity and brain development (Wang et al., 2004; Wang and Zhang, 2012). Enhancer annotations in my dataset contains two novel SEs in the vicinity of *Adcy1* in mouse cerebellum: (1) ∼2.7 kb upstream of its TSS which span 14.2 kb and contain 25 constituent enhancers, and (2) ∼11.5 kb downstream of its TSS which spreads over 81 kb within its gene body and is composed of 47 constituent enhancers (distance calculated from the nearest end of the SEs) (Fig. 4.9A). The constituent enhancers forming the *Adcy1* SEs are highly cerebellum-specific and notably, *Adcy1* is observed to be highly and distinctly expressed in the mouse cerebellum (Fig. 4.9B). Interestingly, the phenotypes associated with the gene knockout of *Adcy1* correlates with the enhancer annotations; the

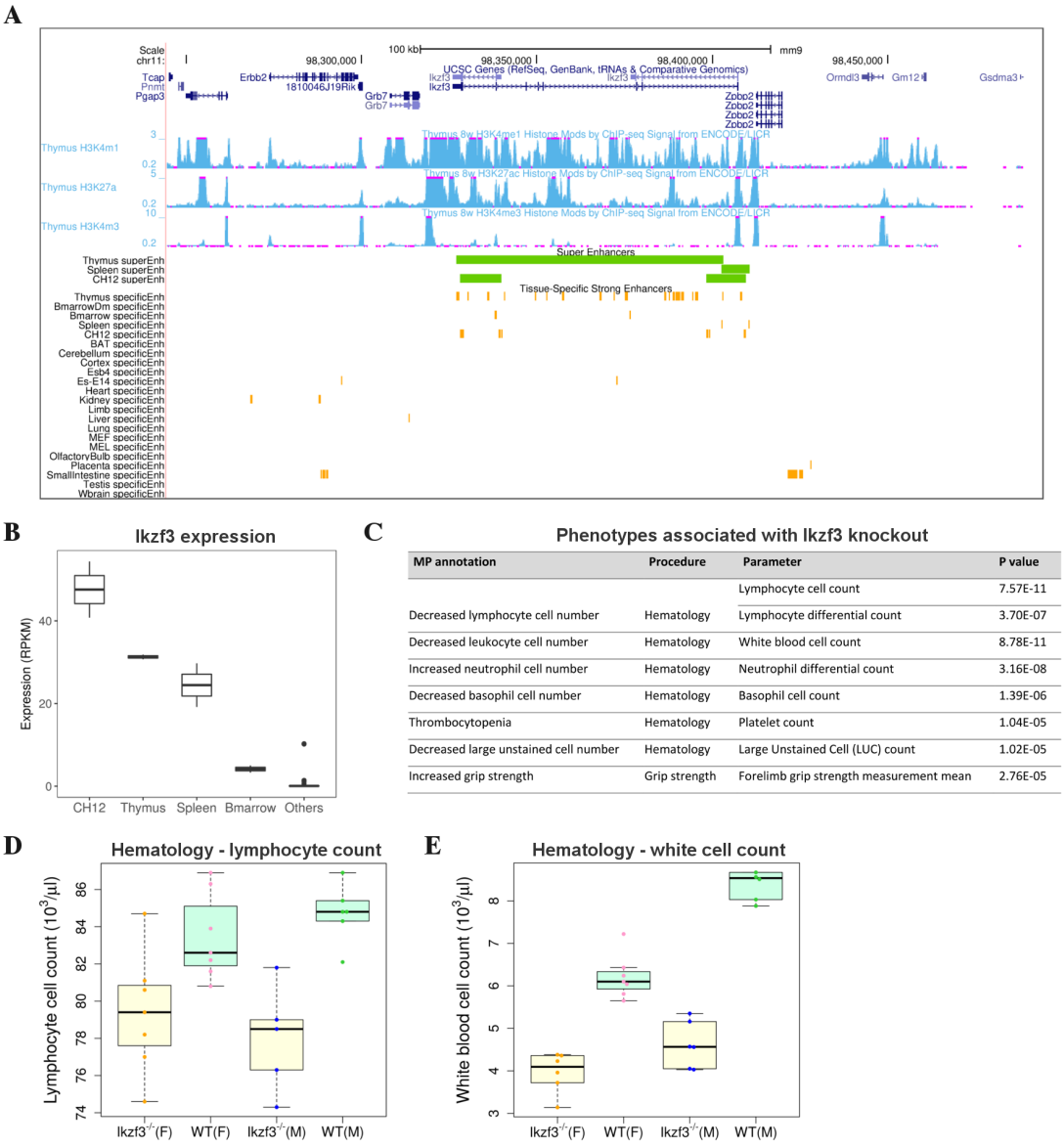


MP annotation	Procedure	Parameter	P value
Abnormal behavior	Open field	Periphery average speed	5.45E-10
		Number of center entries	3.06E-05
Hyperactivity	Open field	Whole arena average speed	5.92E-10
		Whole arena resting time	1.64E-08
		Periphery distance travelled	9.10E-09
		Periphery resting time	1.08E-05
		Distance travelled - total	5.87E-10
Decreased total body fat amount	Body composition (DEXA lean/fat)	Fat mass	1.64E-05
Improved glucose tolerance	Glucose tolerance test	Area under glucose response curve	9.91E-5

Fig. 4.9 Epigenomic landscape and phenotype associations of *Adcy1*. (A) Genome browser snapshot of *Adcy1* locus displaying enhancer profiles in various tissues and cell lines, along with ChIP-seq binding profiles of H3K4me1, H3K4me3 and H3K27ac in the cerebellum. Enhancer annotations: ■ tissue-specific enhancers; ■ super-enhancers. Tissues with no tissue-specific enhancers in this genomic window are not shown. (B) Box plot showing the expression (in RPKM) of gene *Adcy1* across 22 tissues. *Adcy1* is observed to be highly and specifically expressed in the cerebellum. (C) Summary of gene-phenotype associations of *Adcy1* knockout mouse model from the IMPC. The columns show (from left to right) the mammalian phenotype annotation assigned; name of the phenotyping test; parameter of the phenotyping test; p-value of the statistical test performed to compare mutant mice data with the wild-type controls. (D) Box plot comparing periphery average speed between the homozygous *Adcy1* knockout mice and the wild-type controls indicating hyperactive nature of $Adcy1^{-/-}$ mice. (E) Box plot comparing whole arena resting time between the homozygous *Adcy1* knockout mice and the wild-type controls indicating abnormal behaviour of $Adcy1^{-/-}$ mice. For both periphery average speed and whole arena resting time: female (F) control (WT), n = 1753; female homozygous, n = 8; male (M) control (WT), n = 1791; male homozygous, n = 8; linear mixed-effects model was used for the statistical test.

Adcy1 knockout mouse line (*Adcy1*$^{tm1b/tm1b}$) generated by the IMPC shows significant behavioural/neurological associated phenotypes (summarised in Fig. 4.9C). For instance, the *Adcy1*$^{tm1b/tm1b}$ mice exhibit a hyperactive and abnormal behaviour compared to the wild-type controls (Fig. 4.9D-E). Moreover, recent GWAS reports identified two moderately significant SNPs mapped to *Adcy1*; an intron variant ($p = 9 \times 10^{-6}$, rs1521470) in a schizophrenia study (Goes et al., 2015; Welter et al., 2014) and a downstream variant ($p = 5 \times 10^{-7}$, rs116927879) in a bipolar disorder study (Douglas et al., 2016; Le-Niculescu et al., 2009). These observations provide evidence for the potential involvement of cerebellum-specific enhancers in *Adcy1* regulation.

The second example describes the gene Ikaros family zinc finger 3 (*Ikzf3*). *Ikzf3*, a member of the Ikaros group of zinc finger proteins, is a TF which regulates B-cell differentiation and proliferation, and is also important for maturation of T-cells (Morgan et al., 1997; Wang et al., 1998; Winandy et al., 1995). The mouse enhancer annotations in my dataset show that multiple SEs associated with immune system related tissues such as thymus, CH12 and spleen are present near and within the *Ikzf3* gene (Fig. 4.10A). The impact of these enhancers is accordingly reflected in its expression pattern as *Ikzf3* is highly expressed in thymus, CH12 and spleen relative to other tissues (Fig. 4.10B). Consistent with the mouse enhancer annotations, *Ikzf3* is mostly associated with immunological disorders in human and mice. Up-regulation of *Ikzf3* expression has been related with various lymphomas. Billot et al. (2011) observed elevated promoter activity of *Ikzf3* (quantified using H3K4me3) in chronic lymphocytic leukaemia, suggesting epigenetic modifications to play a role in its altered expression. In another investigation, a *Ikzf3* null mutation in mice caused an increase in B-cell precursors and prompted the development of B-cell lymphomas in ageing mutants (Wang et al., 1998). Moreover, *Ikzf3* lacking mice develop phenotypes similar to systemic lupus erythematosus in humans (Sun et al., 2003). Not surprisingly, *Ikzf3* knockout mice characterised within the IMPC also displays predominant immunological phenotypes (Fig. 4.10D-E). In humans, several SNPs mapped to *Ikzf3* in GWASs have been associated to multiple immunological traits including systemic lupus erythematosus, inflammatory bowel disease, asthma and rheumatoid arthritis. The majority of these SNPs are non-coding, mostly occurring within the introns of *Ikzf3*. Overall, these observations demonstrate that gene function is governed by tissue-specific enhancers. However, although genes within the SEC and TEC show significant enrichment of tissue-related phenotypes compared to the background, the enhancer-phenotype relationship is not observed for all the genes. This could be attributed to methodological limitations such as limited number of tissues in my dataset, finite knowledge about gene-phenotype associations, potentially missed/false-positives enhancers, or other genetic and environmental factors affecting the gene function.



MP annotation	Procedure	Parameter	P value
		Lymphocyte cell count	7.57E-11
Decreased lymphocyte cell number	Hematology	Lymphocyte differential count	3.70E-07
Decreased leukocyte cell number	Hematology	White blood cell count	8.78E-11
Increased neutrophil cell number	Hematology	Neutrophil differential count	3.16E-08
Decreased basophil cell number	Hematology	Basophil cell count	1.39E-06
Thrombocytopenia	Hematology	Platelet count	1.04E-05
Decreased large unstained cell number	Hematology	Large Unstained Cell (LUC) count	1.02E-05
Increased grip strength	Grip strength	Forelimb grip strength measurement mean	2.76E-05

Fig. 4.10 Epigenomic landscape and phenotype associations of *Ikzf3*. (A) Genome browser snapshot of *Ikzf3* locus displaying enhancer profiles along with ChIP-seq binding profiles of H3K4me1, H3K4me3 and H3K27ac. Enhancer annotations: ■ tissue-specific enhancers; ■ super-enhancers. (B) Box plot showing the expression (in RPKM) of gene *Ikzf3* across 22 tissues. *Ikzf3* is observed to be highly and specifically expressed in CH12, spleen and thymus. (C) Summary of gene-phenotype associations of *Ikzf3* knockout mouse model from the IMPC. The columns show (from left to right) the mammalian phenotype annotation assigned; name of the phenotyping test; parameter of the phenotyping test; p-value of the statistical test performed to compare the mutant mice data with the wild-type controls. (D-E) The homozygous *Ikzf3* knockout mice predominantly show immunological phenotypes like decreased lymphocyte and white blood cell numbers. Female control (WT), n = 7; female homozygous, n = 6; male control (WT), n = 7; male homozygous, n = 6; linear mixed-effects model was used for the statistical test.

Severity and breadth of phenotypes associated with enhancer target genes

Investigating the phenotypic associations of SEC and TEC showed that both classes are significantly enriched for the corresponding tissue-related phenotypes. However, there is a marked difference in the expression patterns of the SEC compared to TEC (section 4.2.2), which is not observed in their relationship with phenotypes. This dichotomy was further explored by comparing the phenotyping data from knockout mouse lines of genes in SEC and TEC across all tissues within the IMPC data. I reasoned that if SE associated genes are predominantly related to phenotype occurrence, their associated gene knockouts would cause a more severe phenotype condition (a phenotype with an increased effect size) relative to knockouts of other genes (such as those associated with TEs). I compared several standardised phenotyping procedures within the IMPC and observed a significant difference in phenotype severity only for acoustic startle and pre-pulse inhibition (ES $= -0.63$, p $= 0.001$) (Fig. 4.11). However, for the majority of procedures, no significant difference in severity of phenotypes was observed between the SEC and TEC (Open field test: ES $= 0.19$, p $= 0.13$; Grip strength: ES $= 0.19$, p $= 0.55$; Dual Energy X-ray Absorptiometry (DEXA): ES $= -0.02$, p $= 0.75$; Heart weight: ES $= 0.16$, p $= 0.63$; Hematology: ES $= 0.16$, p $= 0.1$) (Fig. 4.11). Next, I sought to examine the breadth of the phenotypes associated with the SEC and TEC. For this purpose, I computed the number of top-level phenotype ontology terms associated with SE and TE associated gene knockouts from the IMPC (Fig. 4.12). No notable difference was observed in the breadth of phenotypes between the SEC and TEC (ES $= 0$, p $= 0.42$), indicating that both SE and TE associated gene knockouts are likely to produce comparable number of phenotypes and therefore, have similar pleiotropic effects. Furthermore, I explored the mouse essential genes by retrieving all the genes from IMPC which generate a lethal knockout (Dickinson et al., 2016) to examine if the SEC is enriched with lethality. There was no significant enrichment of lethal genes amongst the SEC (p $= 0.24$, OR $= 1.08$, 95% CI [0.88, 1.30]) and TEC (p $= 0.83$, OR $= 0.93$, 95% CI [0.79, 1.09]). Overall these results highlight that tissue- and cell-specific relevant traits are associated with both SEC and TEC.

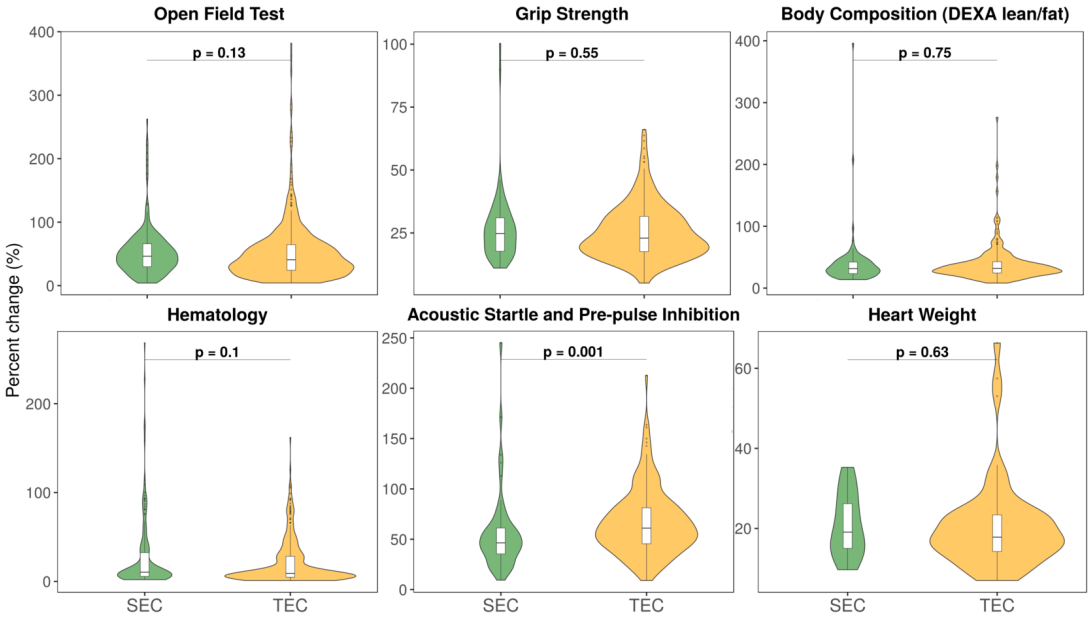


Fig. 4.11 Phenotype severity of SE and TE associated gene knockouts. Violoin plots displaying the percentage change (normalised effect size) in phenotype procedures measured between enhancer associated gene knockouts and wild-type controls. The area under the violin is proportionate to the number of data points in each category. The p-values were calculated using the Wilcoxon Rank Sum Test.

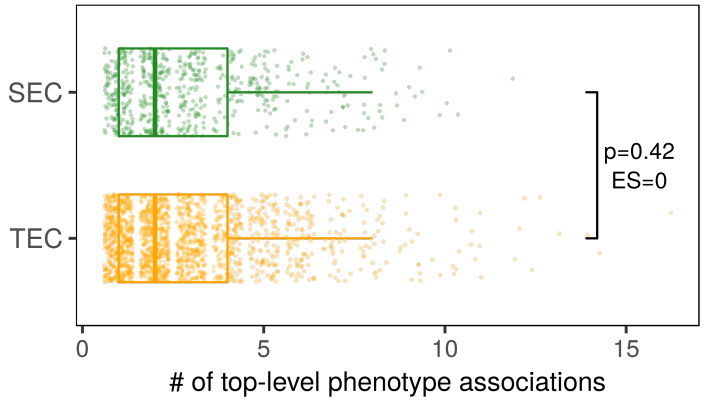


Fig. 4.12 Breadth of phenotypes associated with SE and TE gene knockouts in the mouse. Box plot displaying the number of top-level phenotype terms associated with SE and TE gene knockouts in IMPC. The p-values were computed using Wilcoxon Rank Sum Test. p: p-value; ES: effect size

4.2.4 Protein-protein interactions amongst enhancer associated genes

Having shown that enhancer associated genes are enriched for tissue-specific traits, I hypothesised that the proportion of these with no prior phenotypic annotations related to the tissue maybe involved in disease-causing pathways. These genes are particularly interesting as they might serve as novel candidates for the corresponding tissue-type phenotypes or diseases. Most genes work together with other genes to perform their biological function, thus forming molecular networks where the proteins encoded by the genes interact with each other. These protein-protein interactions (PPIs) are commonly used to examine the relationships between associated genes, based on the well established fact that proteins interacting in a network are likely to be involved in similar metabolic pathways, signalling pathways and cellular processes (Gonzalez and Kann, 2012). Although PPIs are incomplete and susceptible to errors (De Las Rivas and Fontanillo, 2010), several studies have confirmed their usefulness in discovering novel protein function (Sharan et al., 2007) and in identifying functional modules (Dittrich et al., 2008; Spirin and Mirny, 2003). In an event where the normal state of a cell is altered as a result of environmental factors or a disease condition, the PPI networks are also affected (Safari-Alighiarloo et al., 2014). Thus, highlighting the proteins encoded by disease-causing genes in a PPI network can identify other disease-associated risk targets for therapeutic purposes. Therefore, I investigated the PPIs amongst enhancer-associated genes in each of the 22 tissues.

In order to examine the PPIs, I extracted the potential protein interactions from the STRING database (Franceschini et al., 2013) with the highest confidence score (STRING combined score > 0.9). Then in each network, I identified the genes currently known to be associated with the corresponding tissue-type phenotypic annotations from the MGD, while the genes with no-priori phenotypic information were labelled as 'novel'. For each tissue, both the known and unknown disease genes (referred to as known and novel respectively) in the PPI network of enhancer associated genes are observed to be connected in a remarkably dense interactome (Fig.4.13, Appendix A.8). Interestingly, the novel genes (blue nodes) are highly connected with the phenotype-associated genes (pink nodes), suggesting a potential functional relationship between them. Simulating these PPI networks with random protein-coding genes show that novel genes connect significantly more with known phenotype-associated genes, compared to randomly added genes ($p \leq 0.016$, except thymus $p = 0.056$) (Appendix A.9). This outcome demonstrates that enhancer associated genes are potentially engaged in the same functional pathway as the known phenotype genes and therefore, could also contribute to the corresponding phenotypes and ultimately disease causation.

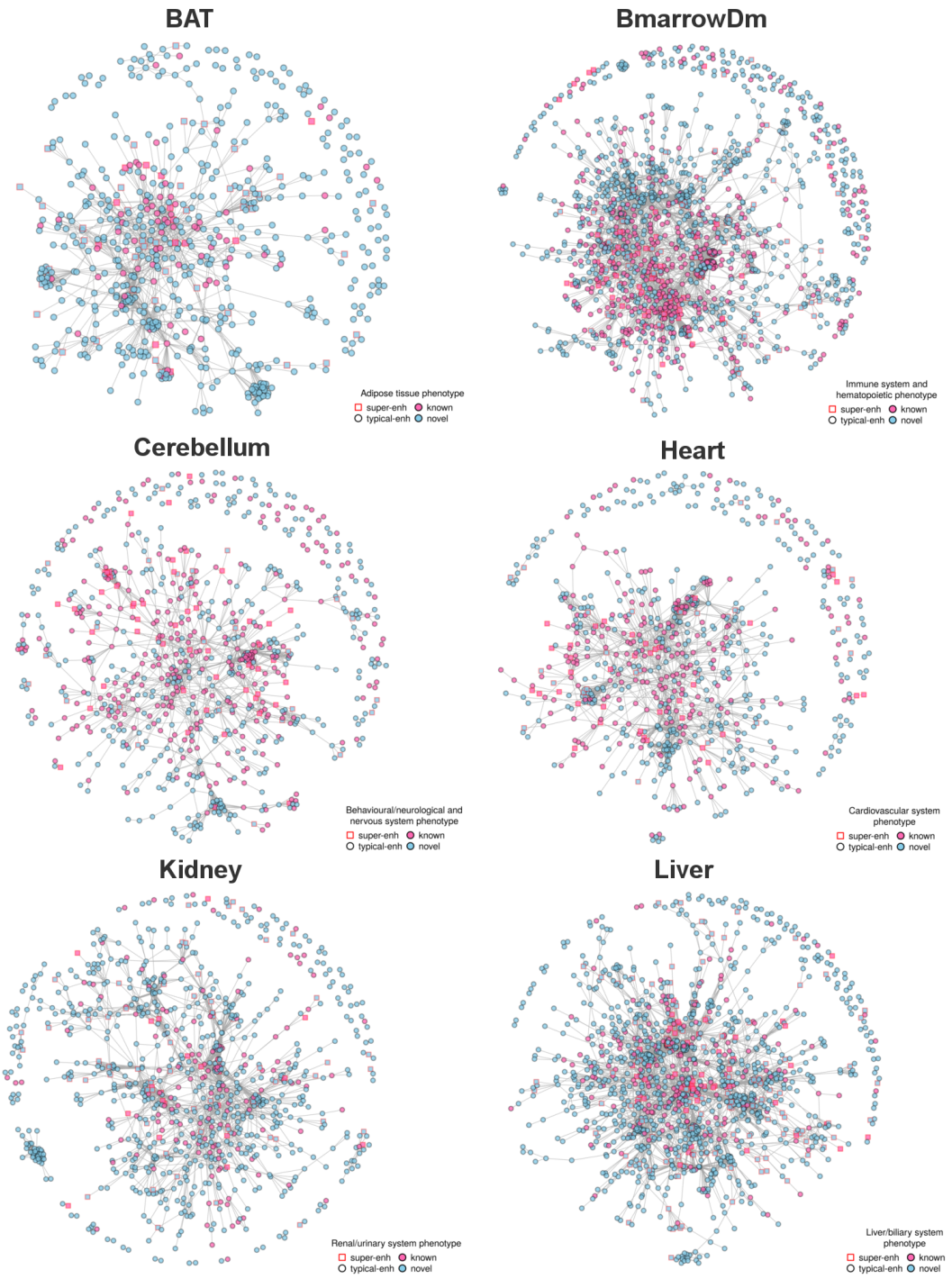


Fig. 4.13 PPI maps of enhancer associated genes. Networks displaying PPIs amongst enhancer associated genes. Nodes in each network represent enhancer associated genes and edges represent potential PPIs. Genes associated with tissue-type relevant phenotypes are highlighted in pink and the shape of the node displays SE and TE associated genes (squares: SEC, circles: TEC). See also Appendix A.8 and A.9.

4.2.5 Transcription factor binding in SEs and TEs

Enhancer regions facilitates binding sites for TFs which contribute to important tissue-specific functions by regulating the target genes (Ong and Corces, 2011). These TFs bind within enhancer regions and recruit additional co-factors to control the expression of their target genes. To investigate TF binding activity within SEs and TEs, with the aim of identifying potential key regulators in each tissue, the Makeev lab used publicly accessible ChIP-seq data for mouse TFs. For many TFs, the information available on their specific binding in various cell-types is rather sporadic, thus the Makeev lab flattened all available ChIP-seq peaks for each TF into single binding profiles referred to as'cistrome' (see methods 4.3.6). Next, for each cell-type, the Makeev lab systematically identified TFs, for which cistrome peaks significantly colocalised with their corresponding active enhancers (see methods 4.3.7).

First, we found that TFs which have significant colocalisation with enhancers include regulators known to be implicated in the corresponding tissue-specific regulation. For example, *Spi1*, with cistrome peaks colocalised with bone marrow enhancers, is implicated in myeloid and B-lymphoid cell development (Scott et al., 1994b); *Gata4*, with cistrome peaks colocalised with heart enhancers, is involved in myocardial differentiation and function (Pikkarainen et al., 2004); and *Dmrt1*, with cistrome peaks colocalised with testis enhancers, plays a key role in male sex determination and differentiation (Raymond et al., 2000). Overall, we observed cistrome peaks of 214 TFs (509 TF-tissue pairs) to significantly colocalise with TEs (with OR > 1; corrected p-value < 10^{-3}) and 113 TFs (148 TF-tissue pairs) with SEs across all tissues and cell-types (Fig. 4.14 shows the top three TFs in each tissue). The 214 TFs colocalised with TEs included all the 113 TFs identified for SEs. Second, we observed that some TFs with cistrome peaks significantly colocalised with enhancers are expressed in a tissue-specific manner in the corresponding tissues (Fig. 4.15A). In total, we identified 56 such TFs with highly tissue-specific expression ($\tau_{exp-frac} \geq 0.85$) and significant colocalisation with corresponding TEs, and 29 TFs with SEs across all tissues and cell-types. Examples of such TFs include *Hnf6* in liver ($\tau_{exp-frac} = 1$), *Nkx2-5* in heart ($\tau_{exp-frac} = 1$), *Gata1* in MEL cells ($\tau_{exp-frac} = 0.93$) and *Neurog2* in whole brain ($\tau_{exp-frac} = 0.98$).

Overall, TF cistrome peaks were identified to significantly colocalise with both SEs and TEs, but a greater number of TFs were identified to colocalise with TEs compared to SEs. This could be explained by the relatively large number of TEs in the genome. To investigate this further, for each TF with significant enhancer localisation, the Makeev lab computed their TFBS density in SEs and TEs. The TFBS density could be defined as a measure of TFBS clustering in SEs or TEs (see methods 4.3.8). To summarise this analysis, the Makeev lab counted the number of TF-tissue pairs which have significantly

greater TFBS density in SEs compared to TEs, and vice-versa for TEs. Overall, we find that SEs have more TF-tissue pairs with higher TFBS density compared to TEs (Fig. 4.15B). Altogether, this data indicates that although TEs are more often colocalised by TF cistrome peaks, frequency and degree of TFBS clusters is higher in SEs.

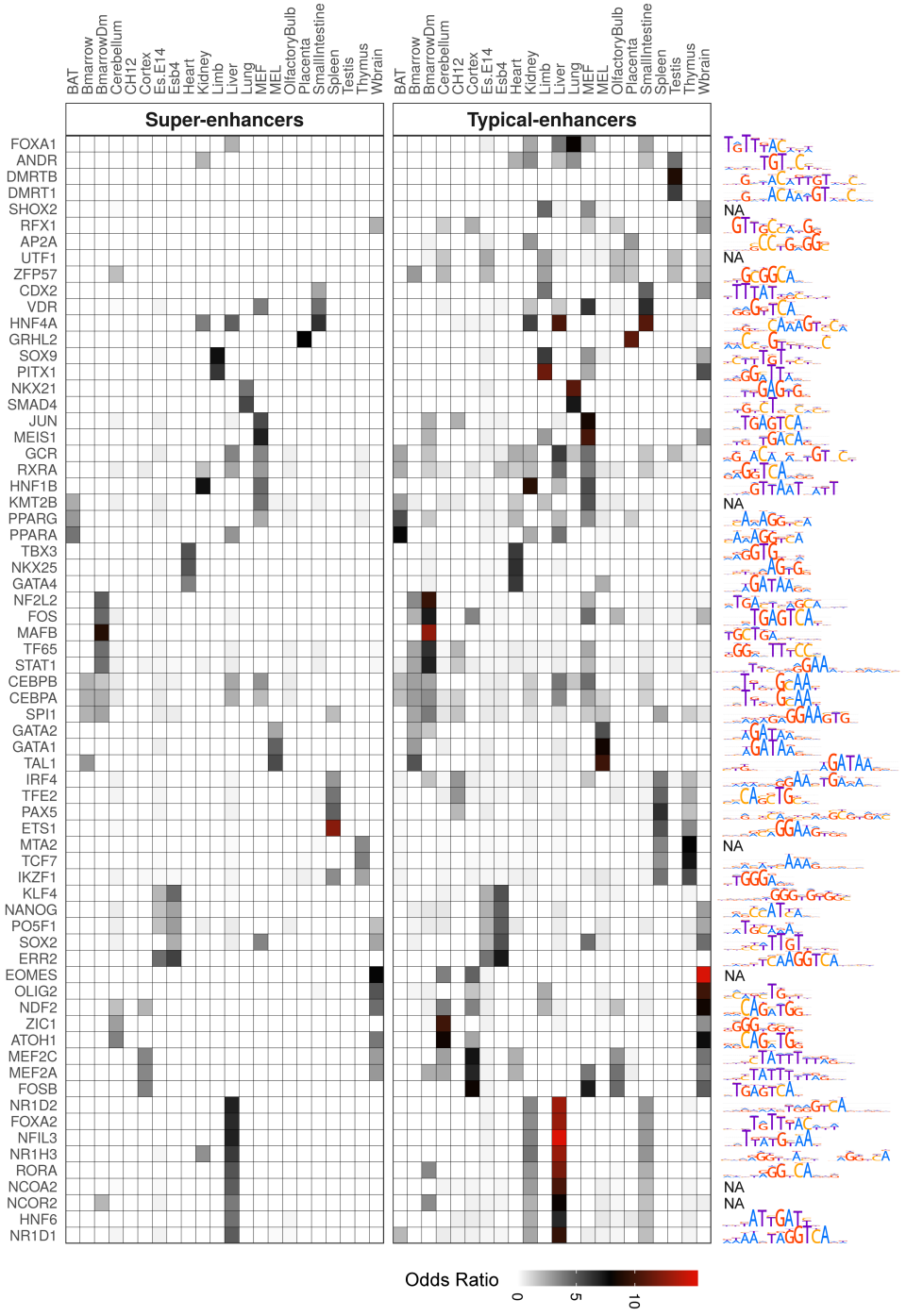


Fig. 4.14 Master regulators enriched in SEs and TEs. Heatmap showing the top three enriched TFs identified in SE and TE constituents in each tissue. The motifs associated with the enriched TFs are shown on the right. NA is shown for TFs with motifs not present in HOCOMOCO v11 (Vorontsov et al., 2018). The rows of the heatmap are clustered using hierarchal clustering.

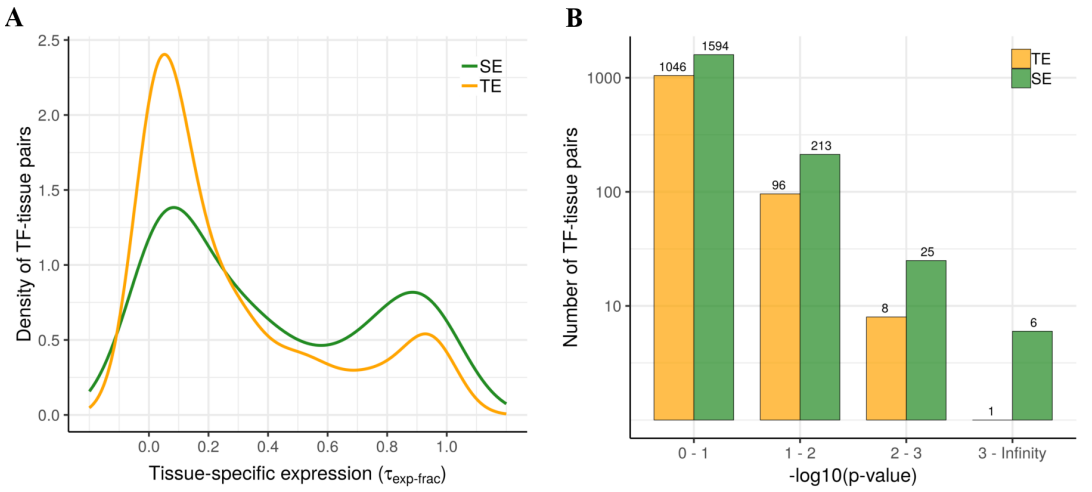


Fig. 4.15 Transcription factor binding within SE and TE constituents. (A) Density plot showing the distribution of TFs whose cistrome significantly colocalised with enhancer segments, plotted against the tissue-specific expression of the TF in the corresponding tissues. (B) Bar plot displaying the number of TF-tissue pairs which have significantly greater TFBS density in SE compared to TEs (green bars), and vice-versa for TEs (orange bars). For each TF-tissue pair, a Wilcoxon Rank Sum Test was used to compare its TFBS density between SEs and TEs. The TF-tissue pairs are binned by the logarithmic significance of the difference in TFBS density between SEs and TEs obtained from the Wilcoxon Rank Sum Test. For e.g. the first pair of bars represents the TF-tissue pairs with a p-value of difference lying in the range $0 \leq -log_{10}(p-value) < 1$.

4.2.6 Combinatorial learning approach for phenotype prediction

Up until now, the analysis presented here shows that the mouse enhancer regions are correlated to a great extent with total-expression, tissue-specific expression and phenotypes. While PPI and gene expression data has been previously used to infer mammalian function (Pena-Castillo et al., 2008; Tasan et al., 2008; Yuan et al., 2012), I sought to evaluate the capability of predicting gene-phenotype associations by utilising regulatory data predicted here such as TSREs and TF binding. I implemented the random forest algorithm to predict gene-phenotype associations from 13 different phenotypic domains, where each domain is relevant to at least one tissue type in my dataset. For this approach, I extracted gene features from TSRE profiles, expression data, TFBSs and PPI data in 22 mouse tissues (see methods 4.3.9) (features summarised in Table 4.1). To make gene-phenotype predictions, a random forest classifier was first trained on a subset of protein-coding genes using a combination of various gene features as predictor variables and the top level mammalian phenotype terms from the MGD as the response variable (true positives), while genes not associated with a phenotype in the MGD were considered as true negatives. This model was used to predict gene-phenotype associations in the remaining set of genes not used in the training of the model.

Table 4.1 Summary of the gene features used in the random forest classifier.

Data	Feature	Description	n	Symbol
Regulatory elements	Tissue-specific enhancer profiles	Sum of posterior probabilities for all tissue-specific strong enhancers associated in each tissue	22	TSRE
	Tissue-specific promoter profiles	Sum of posterior probabilities for all tissue-specific active promoters associated in each tissue	22	
	Transcription factor binding	Enrichment of motifs within cistrome regions overlapping 500 bp upstream and 100 bp downstream of TSS	297	TF
Protein-protein interactions (PPI)	PPI with genes associated with enhancers	PPI score of a gene within tissue-specific enhancer network	22	TSRE_PPI
	PPI with genes associated with promoters	PPI score of a gene within tissue-specific promoter network	22	
	PPI with genes associated with the phenotype	PPI score of a gene within a phenotype associated network	1	PPI
Expression	Expression profiles	Expression of a gene in each tissue	22	Exp

By integrating various features together, 10 combinations were formed, constructing 10 distinct classifiers for each phenotypic domain. The predictive power of each classifier was assessed by generating Receiver Operating Characteristic (ROC) and precision-recall (PR) curves based on 5-fold cross validation repeated 10 times with different seeds. The random forest classifiers trained on just the regulatory elements (enhancers, promoters and TFs) achieved the poorest performance compared to other models with a mean AUC-ROC (area under the ROC curve) of 0.71 and AUC-PR (area under the PR curve) of 0.19 across all the phenotypes (Fig. 4.16A). Conversely, the classifier trained on all the gene features combined (Exp+PPI+TSRE+TSRE_PPI+TF) achieved the best performance with a mean AUC-ROC of 0.78 and AUC-PR of 0.27 across all the phenotype domains (Fig. 4.16A). For all phenotypes, the mean AUC-ROC for this model exceeded 0.74 and AUC-PR ranges between 0.11 and 0.43. However, high precision recall rate (AUC-PR $>$ 0.35) was observed in phenotypes with a high number of known mammalian phenotype annotation counts in the MGD (such as behavioural/neurological, nervous system, cardiovascular, immune and hematopoietic system) (Fig. 4.16B), indicating that precision in predicting gene-phenotype associations is dependent on the amount of true positives used to train the classifier.

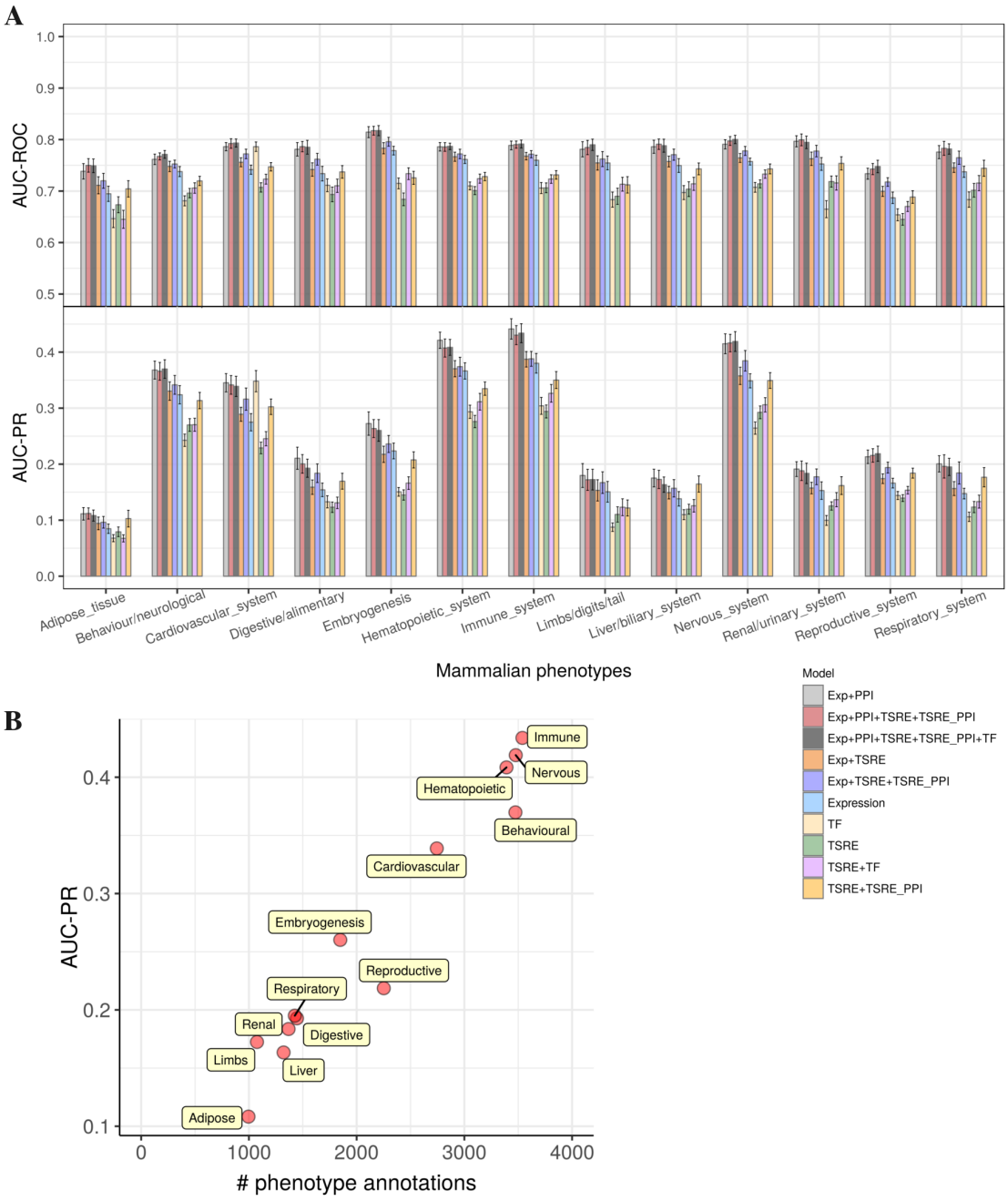


Fig. 4.16 Evaluation of classifiers to predict gene-phenotype associations in the mouse. (A) Bar plots comparing the predictive power of various random forest classifiers across various phenotypes. Top panel displays the area under the curve (AUC) measured for ROC curves while bottom panel displays AUC for PR curves. The range of y-axis has been adjusted to clearly show the differences between the various models. Error bars denote the standard deviation. (B) Scatter plot showing the relationship between the precision rate and the number of known phenotype annotation counts in the MGD for each phenotype domain.

As an example, I describe here the results of predicting gene-phenotype associations within the nervous system domain. For nervous system phenotype, the classifier trained on all the gene features achieved the greatest mean AUC-ROC of 0.80 and AUC-PR of 0.42 (Fig. 4.17A-B). From the cross validation results of this classifier (5-fold cross validation repeated 10 times i.e. 50 runs in total), one can clearly observe the modelled classifier to be robust with a low standard deviation (Fig. 4.17C-D). The PPI score with genes known to be associated with nervous system phenotype was identified to contribute the most in predicting nervous system gene-phenotype associations, followed by expression data in whole brain and cortex (Fig. 4.17E). In fact, PPI data was the most informative and the main contributor to the performance of these classifiers in all the 13 phenotypes (Appendix A.10). The classifier trained on only expression and PPI data also performed well (AUC-ROC = 0.79) and the predictions were comparable to the best performing model. While the models trained solely on regulatory features had limited predictive power, they improved the performance of models when integrated with other features, suggesting that regulatory data are a useful addition for modelling mammalian phenotypes.

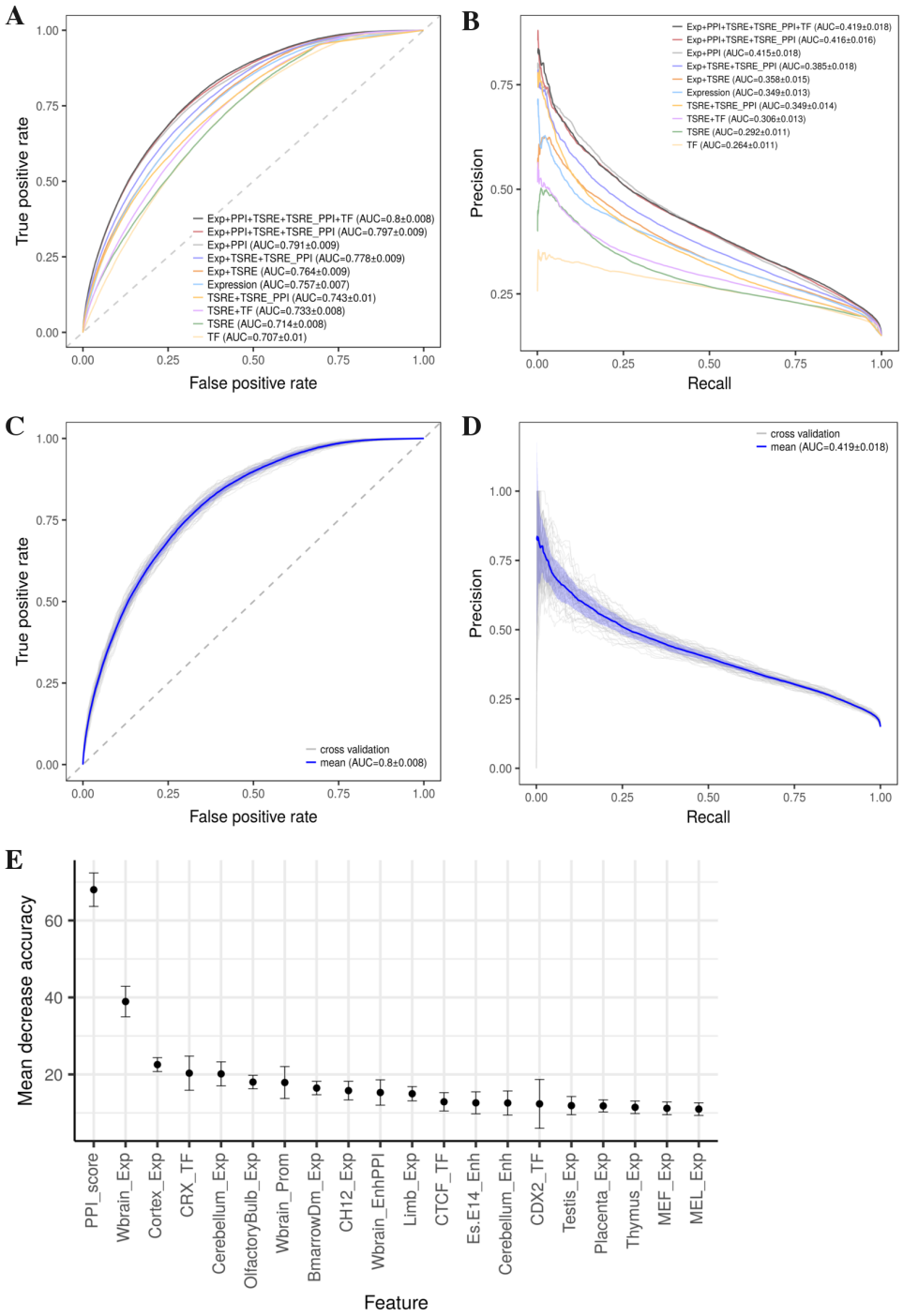


Fig. 4.17 Predicting genes associated with nervous system phenotype. (A) ROC and (B) PR curves comparing the performance of 10 random forest classifiers modelled to predict genes associated with nervous system phenotype. ROC and PR of various models was measured by calculating the AUCs. (C) ROC and (D) PR curves displaying the cross validation results for the best performing model (Exp+PPI+TSRE+TSRE_PPI+TF). Grey lines represent the cross validation runs, blue lines represent the mean of the cross validation and the blue shaded area shows the standard deviation. (E) Feature importance chart of the best performing model showing the top 20 predictor variables important in contributing to the nervous system phenotype predictions, as measured by the mean decrease in accuracy (x-axis) determined by the random forest model. Exp: expression; Enh: enhancer; Prom: promoter; TF: transcription factor. See also Appendix A.10.

Next, I investigated the novel gene-phenotype predictions made by these classifiers. The predictions from these classifiers were evaluated based on the current knowledge of gene-phenotype associations. However, there may be cases where there is no, or little prior knowledge about the function of a gene. For such cases, the predictions from the classifier would be categorised as incorrect, but it is possible that these associations could be novel. These cases also leads to undermining of the true predictive power of a classification model. For such reasons, the top false-positive predictions are the most interesting as they could provide new hypotheses about gene function. In order to capture all the novel predictions, I trained the classifiers on all the protein-coding genes and extracted top scoring false-positives with a prediction probability ≥ 0.90. As an example case study, I further investigated the top novel predictions for the nervous system phenotype. For the 76 false-positive predictions (prediction probability ≥ 0.90) identified for nervous system domain, I examined their PPIs with genes associated with Alzheimer's disease (AD) - which is the most enriched neurodegenerative disorder amongst the mouse genes currently annotated with nervous system phenotype (Appendix B.5). Out of the 76 predicted novel genes, 42 were connected to known AD genes while 34 genes had no available PPI data. Eleven genes (*Wnt3, Vip, Adcy2, Esrrg, Htr5a, Apln, Lpar4, Pnoc, Sstr3, Htr6, Calcr*) stood out in this network as they were densely connected (≥ 10 interactions) with AD associated genes while a further 31 had at least one interaction (Fig. 4.18). Interestingly, out of these 11 highly connected genes, 8 were associated with G-protein coupled receptor signalling pathway (Table 4.2). Numerous previous studies have documented the role of G-protein coupled receptors in the pathogenesis of AD (Qiu, 2017; Thathiah and De Strooper, 2011; Zhao et al., 2016), which suggests that these novel predictions might serve as useful candidates to investigate the AD functional pathway.

In order to systematically examine the top false-positive predictions (prediction score ≥ 0.90) from all phenotype domains, I used the Open Targets Platform (Koscielny et al., 2017) which links potential novel genes to diseases via evidence based on genetic associations, somatic mutations, animal models, expression, pathways, drugs and text mining from literature. Overall, I identified that $\sim$75% (495/659) of the examined false-positive predictions could be potentially associated with the corresponding disease (Fig. 4.19) and hence, could serve as potential novel disease targets. For instance, out of the 76 top scoring false-positives examined for nervous system phenotype, 72 are likely to be associated with nervous system disease ($p = 5.00 \times 10^{-9}$) based on evidence integrated from a range of data sources. Figure 4.20 displays the top 10 gene predictions associated with the corresponding disease for various phenotype domains and the evidence supporting their association.

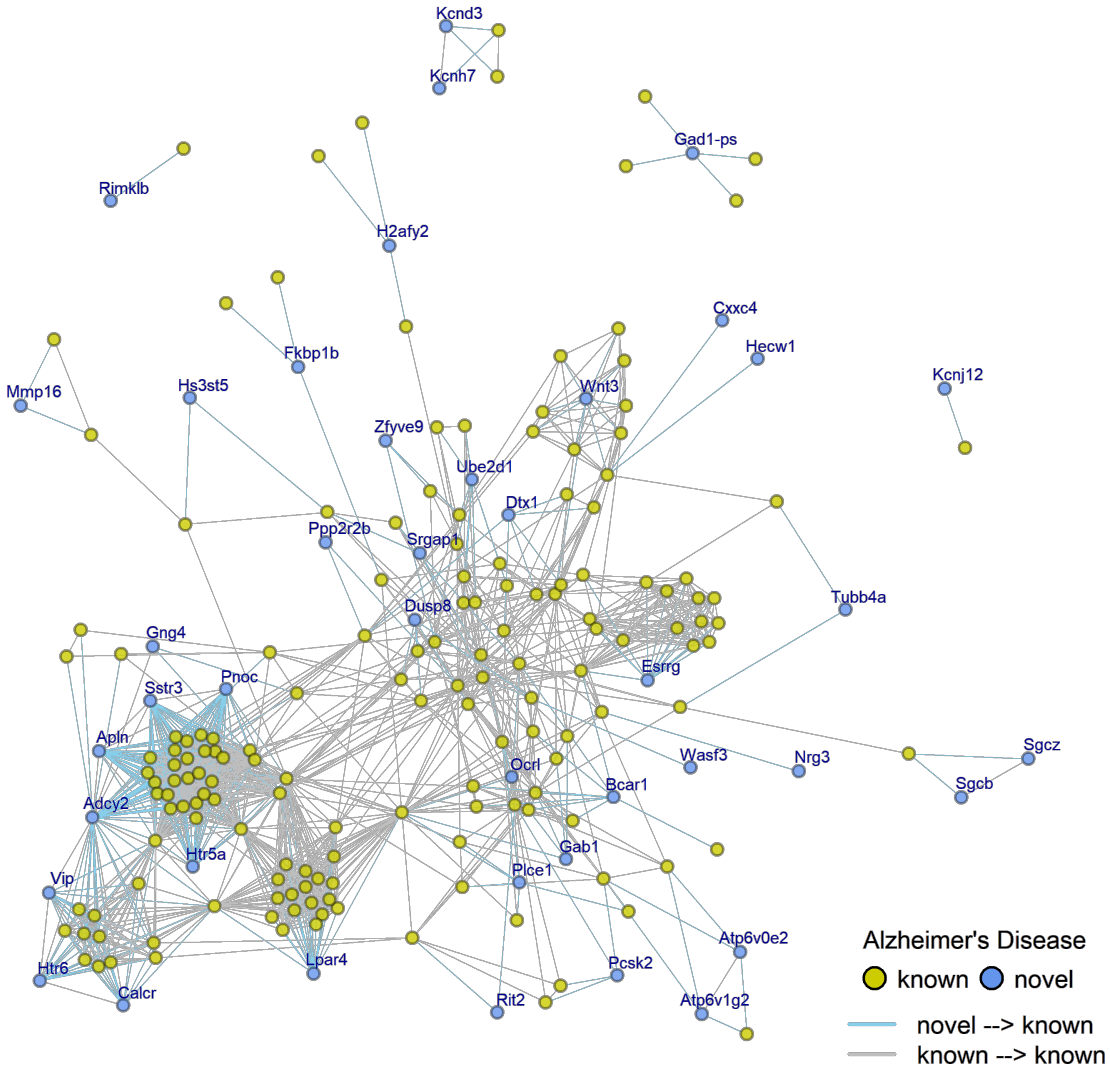


Fig. 4.18 PPI map of novel nervous system phenotype predictions with AD associated genes. Nodes in the network represent genes and edges represent PPIs obtained from STRING. Yellow nodes represent genes known to be associated with AD, while blue nodes display potential novel genes predicted to be associated with nervous system phenotype by the random forest classifier. Grey edges represent PPIs between two known AD associated genes or two novel predicted genes, while PPIs between a known AD gene and a novel gene are highlighted in blue.

Table 4.2 GO enrichment analysis of 11 novel nervous system phenotype predictions identified to be densely connected with AD associated genes.

Term name	Term id	Domain	Query size	Overlap size	Corrected P-value	Genes
G-protein coupled receptor signalling pathway, coupled to cyclic nucleotide second messenger	GO:0007187	BP	11	4	9.47E-04	*Vip, Calcr, Htr5a, Sstr3*
signal transduction	GO:0007165	BP	11	11	1.69E-03	*Wnt3, Vip, Adcy2, Calcr, Esrrg, Htr6, Apln, Htr5a, Sstr3, Pnoc, Lpar4*
G-protein coupled receptor signalling pathway	GO:0007186	BP	11	8	2.83E-03	*Vip, Calcr, Htr6, Apln, Htr5a, Sstr3, Pnoc, Lpar4*
signalling	GO:0023052	BP	11	11	3.47E-03	*Wnt3, Vip, Adcy2, Calcr, Esrrg, Htr6, Apln, Htr5a, Sstr3, Pnoc, Lpar4*
cell communication	GO:0007154	BP	11	11	3.81E-03	*Wnt3, Vip, Adcy2, Calcr, Esrrg, Htr6, Apln, Htr5a, Sstr3, Pnoc, Lpar4*
cellular response to stimulus	GO:0051716	BP	11	11	1.33E-02	*Wnt3, Vip, Adcy2, Calcr, Esrrg, Htr6, Apln, Htr5a, Sstr3, Pnoc, Lpar4*
cAMP-mediated signalling	GO:0019933	BP	11	3	1.64E-02	*Vip, Calcr, Htr5a*
cyclic-nucleotide-mediated signalling	GO:0019935	BP	11	3	2.26E-02	*Vip, Calcr, Htr5a*
cAMP biosynthetic process	GO:0006171	BP	11	3	3.54E-02	*Vip, Adcy2, Calcr*

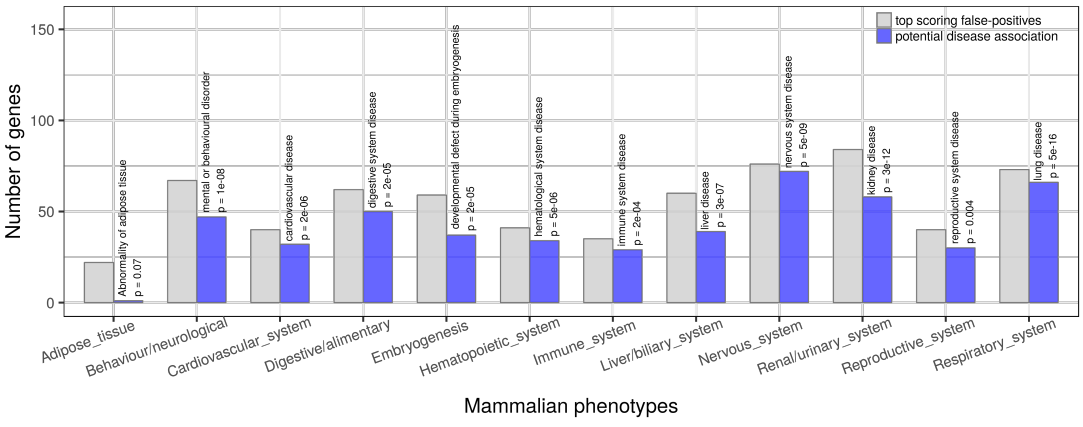


Fig. 4.19 Evaluation of the top scoring false-positives from random forest classifiers. Bar plot displaying the number of genes in each phenotype domain with potential association to the corresponding disease.

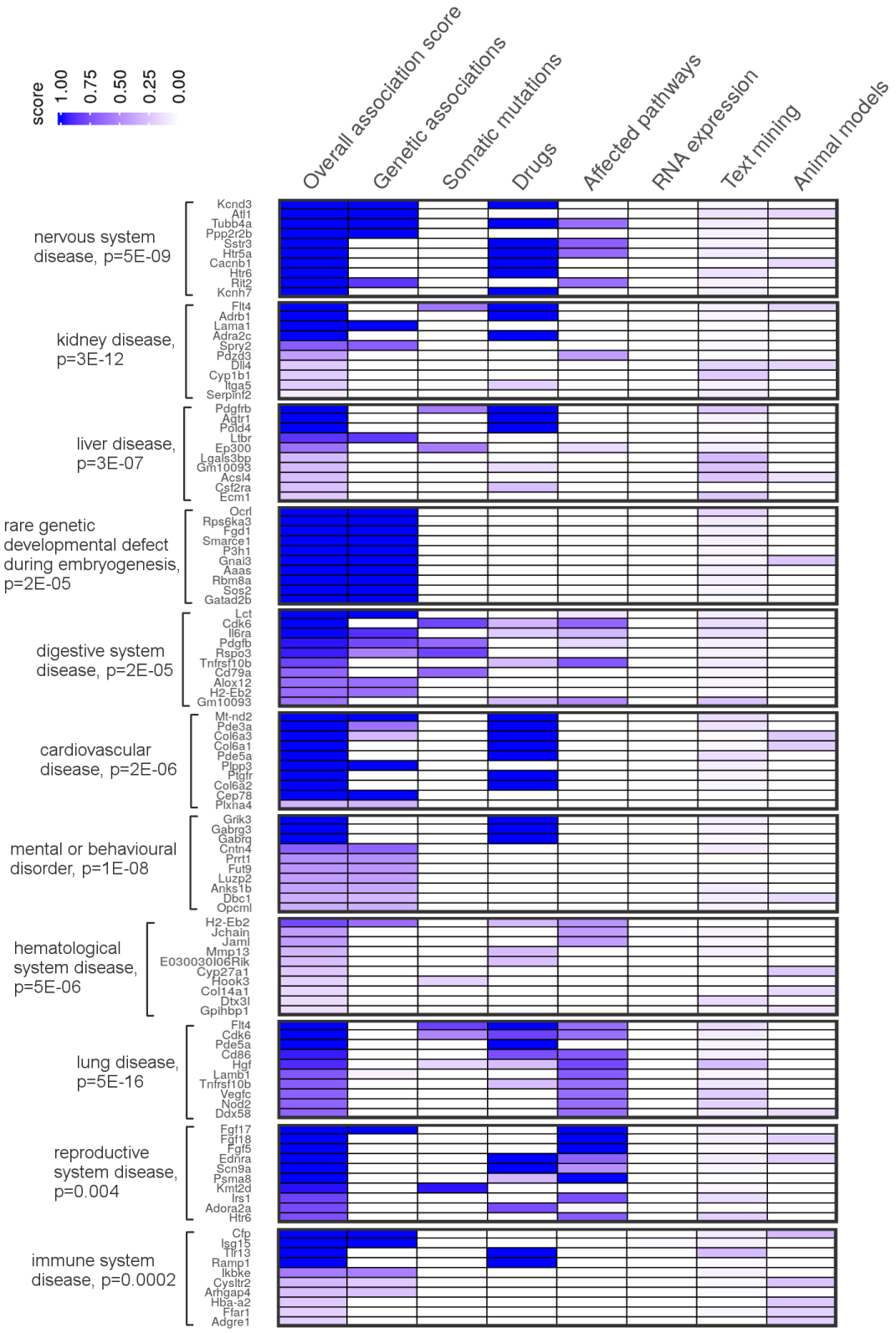


Fig. 4.20 Top scoring novel gene predictions. Heatmap displaying the top 10 gene predictions in various phenotype domains along with the evidence source for their potential association to the corresponding disease. The Open Targets Platform was used to link the novel predictions to diseases using integrated genome-wide public datasets.

4.3 Methods

4.3.1 Datasets

For investigating the expression of enhancer associated genes, RNA-seq data for all 22 tissues and cell lines was collected from ENCODE as read alignments (BAM files). Data for cell lines CH12 and Es-E14 was collected from Standford/Yale lab while rest of the data was retrieved from LICR lab. From the BAM files, the read counts over all genes (mm9, ensembl v67) were quantified using HTSeq (Anders et al., 2015) and expression of each gene was calculated in RPKM in each tissue/cell line. A mean RPKM value was calculated for multiple biological replicates from ENCODE.

4.3.2 Associating TSREs to potential target genes

GREAT (McLean et al., 2010) was used to associate tissue-specific regulatory elements to potential target genes in each tissue. In cases where GREAT predicted multiple target genes for a particular TSRE, the gene nearer to the TSRE was selected as the primary predicted target for all further downstream analysis. GREAT was run using default parameters on mm9 assembly and the whole genome was selected for control background regions. To examine the consistency of predicted enhancer-gene assignments with other datasets, they were compared to previously reported TADs (Dixon et al., 2012) and EPUs (Shen et al., 2012) in the mouse genome. The enhancer-gene pairs across the 22 tissues were merged together for this comparison. The TADs (in mESCs and cortex) were compared to the enhancer-gene pairs to examine if the enhancer-gene pair overlaps the same TAD. Only the cases where both the enhancer and its associated gene overlapped a TAD were used. This analysis identified 96.62% and 93.57% of enhancer-gene pairs to be in the same TAD annotated in cortex and mESCs respectively. A similar comparison was done with EPUs which revealed 87.23% of predicted enhancer-gene pairs to be in the same EPU.

4.3.3 Expression analysis of enhancer associated genes

To examine the relationship between enhancers and expression of their target genes, data from all 22 tissues was combined into gene-tissue pairs and grouped into three classes based on their enhancer association: (1) gene-tissue pairs associated with SEs (SEC), (2) gene-tissue pairs associated with TEs (TEC), and (3) gene-tissue pairs associated with weak enhancers (WEC). In order to quantify tissue-specific expression of target genes, I calculated the tissue-specificity index of each gene using the Tau method described

earlier in chapter three (section 3.3.4). A matrix of expression values was constructed with dimensions $t \times s$, where t is the total number of genes and s is the number of tissues/cell lines. Genes not expressed in any tissue were deleted from the matrix leaving genes expressed in at least one tissue. The RPKM values were log2 transformed and quantile normalised (using the normalize.quantiles function in preprocessCore R package) to allow easier comparison of gene expression across tissues. Genes were then sorted by ascending quantile normalised values and divided into deciles of equal density and placed into 10 bins. The lowest decile (lowest 10% of genes by QN value) was placed in bin 1, the next lowest was placed in bin 2, and so on until the top 10% of quantile normalised values were placed in bin 10. The Tau value (τ_{exp}) for each gene was calculated as:

$$\tau_{exp} = \frac{\sum_{i=1}^{N}(1-\hat{y_i})}{N-1}; \quad \hat{y_i} = \frac{y_i}{max(y_i)} \tag{4.1}$$

where N is the total number of tissues, $\hat{y_i}$ is the normalised expression bin profile component of the gene in tissue i. In order to associate τ_{exp} values to tissues, the Tau-fraction (represented as $\tau_{exp-frac}$) for each gene in every tissue was calculated as $\frac{\tau_{exp} \times y_i}{M}$, where y_i is the expression of the gene in tissue i and M is the maximum expression of the gene across all the tissues. Genes with $\tau_{exp-frac} \geq 0.85$ were categorised as having high tissue-specific expression in the corresponding tissue (Kryuchkova-Mostacci and Robinson-Rechavi, 2017; Yanai et al., 2005). Housekeeping genes were identified based on a strict τ_{exp} threshold. Genes with low τ_{exp} score (≤ 0.20) are uniformly expressed across all the tissues and were considered to be 'housekeeping' genes. A total of 1,252 housekeeping genes were identified using this threshold out of which 1,171 were protein-coding genes.

To visualise the distinct number of enhancer tissue-types calculated for each enhancer-associated gene (Fig. 4.5B), binary matrices for SEC and TEC in 22 tissues were generated such that the rows in the matrix represent enhancer associated genes and column represent different tissues. A value of '1' or '0' was assigned to the cells in the matrix depending on if the gene was identified to be associated with the enhancer of that tissue or not respectively. The heatmaps in Figure 4.5B were first sorted on number of tissue-type associations of genes and then sorted by the order of tissues across the columns.

4.3.4 GO, mammalian phenotype and disease enrichment analysis

To investigate the molecular functions and biological processes linked with enhancer associated genes, I combined the SE and TE associated genes across the 22 tissues to make two unique list. This resulted in 3,617 genes to be associated with only SEs and 11,437 genes to be associated with only TEs. These gene sets were then used

for GO enrichment analysis using the ToppGene suite (Chen et al., 2009) (Appendix B.1, B.2). The enrichment of mammalian phenotypes and human diseases in SEC and TEC was calculated individually in each tissue using the ToppFun tool in ToppGene suite. Fisher's exact test was used for calculating the enrichment of housekeeping genes amongst the SEC and TEC. For background, the total number of protein-coding genes in the genome was used. The SEC was observed to be significantly depleted for housekeeping genes (155/3,617; $p = 0.012$, OR = 0.82), while the TEC was enriched (686/11,437; $p = 2.7 \times 10^{-11}$, OR = 1.49).

The enrichment of enhancer associated genes in mammalian phenotypes was computed using gene-phenotype associations in the MGD, collected on 14th June 2017. The total number of genes associated with a phenotype in the MGD (16,494) were used as the background in this case. For enrichment using the IMPC data, all the statistically significant genotype-phenotype associations and their phenotyping data for IMPC release version 5.0 were collected from the IMPC website. This compromised of phenotype data for 3,323 gene knockouts, with 2,900 genes significantly associated with at least one phenotype attribute ($p \leq 10^{-4}$). The IMPC consists of data from various standardised phenotype procedures whose protocols are described in the IMPReSS database (https://www.mousephenotype.org/impress). For each phenotype procedure, the total number of genes tested for that particular procedure were used as the background for calculating enrichments. To quantify the severity of phenotypes, I used the percentage change value from each procedure. The percentage change is the normalised effect size, which is scaled to make it comparable across various procedures and parameters (Kurbatova et al., 2015). The percentage change between SE and TE associated genes was compared for several standardised phenotyping procedures. All the parameters within a procedure were grouped together for this analysis. For computing the enrichment of mouse essential genes in the SEC and TEC, genes producing a lethal knockout (960 genes out of 2,900) were used.

4.3.5 Protein-protein interaction maps

The predicted PPIs amongst the genes of interest were extracted from the STRING database (Franceschini et al., 2013) using the R package STRINGdb. A score threshold of 900 was implemented to extract potential interactions with the highest confidence and reduce false-positives. The interaction maps were visualised as networks using the iGraph package in R. The known gene-phenotype associations (from MGD) in the network were labelled 'known' while the remaining genes were marked 'novel'. A permutation test was performed to identify if the observed number of interactions between known and novel genes are more than what one would expect by random. I

added randomly selected protein-coding genes equal to the number of genes known to be associated with phenotypes in the network and extracted their interactions from STRING. The number of interactions (edges) between randomly added genes and known phenotype genes were then counted. This was repeated 1,000 times to produce a distribution of expected number of edges and the p-value was calculated as $p = y/N$, where y is number of permuted random-known edges greater than the observed novel-known edges and N is the total number of items in the distribution (i.e. 1,000).

4.3.6 Cistrome data

For the analysis of TFBSs colocalised with different enhancer sets, the Makeev lab used a cell-type independent cistrome, the general genomic map of regions bound by particular TFs in any cell-type (Vorontsov et al., 2018). The cistrome is based on uniformly reprocessed ChIP-seq data from the GTRD database (Yevshin et al., 2018) across all the cell-types and conditions. The cistrome regions were classified into four reproducibility categories (A,B,C,D): A - regions supported by ChIP-seq data from two different experimental data sets (at least one was accompanied by control data) and different ChIP-seq peak calling tools; B - regions supported by peak calls from two different experimental data sets (at least one was accompanied by control data); C - regions supported by peak calls from a single experimental data set with control data and different peak calling tools; D - all other reproducible regions (supported by more than one peak). A and B categories were taken into the analysis by default. For TFs with a limited number of ChIP-seq data sets, the Makeev lab added regions from C and D categories when it was necessary to get at least 100 peaks. As an additional filter for cistrome, the Makeev lab used TF binding motifs from HOCOMOCO (Vorontsov et al., 2018) to annotate motif occurrences in cistrome regions with SPRY-SARUS (Kulakovskiy et al., 2016) using the default motif p-value threshold of 5×10^{-4} (Kulakovskiy et al., 2013) and then discarded cistrome segments without motif occurrences.

4.3.7 Enrichment of TFBSs in SEs and TEs

To calculate the enrichment of TF binding within SE and TE constituents, the Makeev lab first merged the neighbouring constituent enhancers within 400 bp into prolonged extended enhancer segments in each tissue. These extended enhancer segments were then used to generate the control regions; more precisely, for each enhancer segment of length L, the Makeev lab located two segments (enhancer shades) of length L, one at $100 \times L$ upstream and the other at $100 \times L$ downstream. This produced a set of control

segments of the same lengths and similar global genomic context as the enhancer segment under study. The Makeev lab checked if any control segments overlapped other constituent enhancers, but such cases contributed only 1-2% of the total number of control regions and were safely ignored. The extended enhancer segments and control regions were then intersected with the cistrome peaks of each TF and split into two groups; overlapping (if least 1 bp overlapped) and non-overlapping with the cistrome. The Fisher's exact test on 2×2 contingency tables was used to assess the statistical significance of TF cistrome peaks overlapping constituent enhancers (SE or TE) versus control regions. The resulting p-values were corrected for multiple testing using Bonferroni correction. Note that the cistrome segments of a TF can significantly colocalise with enhancers in several different cell-types, therefore, the Makeev lab counted the number of significant enrichments as TF-tissue pairs. The Makeev lab also performed the analysis with only the cistrome segments that contain high scoring motif hits from HOCOMOCO. The results were very similar to the analysis where all cistrome segments were considered; about 10% of TFs did not have known binding motifs, and for TFs with known motifs, about 90% of significant TF-tissue pairs were independent from whether the motifs were considered or not.

4.3.8 TFBS density analysis

To calculate the TFBS density of each TF, the Makeev lab intersected each enhancer with the TF cistrome peaks. Within these overlapping regions, the Makeev lab predicted the binding motif occurrences of the corresponding TF using HOCOMOCO-v11 motifs. In cases where HOCOMOCO contained multiple motif models for a single TF, all motifs were used and the binding sites exceeding the cistrome p-value threshold of 0.0005 were retained. Density was calculated as the total genomic coverage of motifs (in bp) divided by the total coverage of enhancer-cistrome intersection (in bp). The Makeev lab calculated densities for only those enhancers (constituent enhancers of SEs or TEs) which had at least one motif occurrence in its intersection with the cistrome. The Wilcoxon Rank Sum Test was then used to compare the TFBS densities of TF-tissue pairs in SEs and TEs (each TF-tissue pair was compared individually between SEs and TEs). The non-corrected p-values were used to order the TF-tissue pairs by their level of TFBS density disparity between SEs and TEs. The TF-tissue pairs were grouped into bins based on their p-value and the number of TF-tissue cases where its TFBS density was more in SEs compared to TEs, or vice versa, were counted (Fig. 4.15B).

4.3.9 Predicting gene-phenotype associations

To predict mammalian gene-phenotype associations, features were extracted from TSREs, expression, TF binding and PPI data for all protein-coding genes. From the TSRE profiles across 22 tissues, strong-enhancers and active promoters associated with each protein-coding gene were extracted. A score representing the tissue-specific enhancers and promoters in each tissue was computed as $S_{gt} = \sum_{i=1}^{N}(P)$, where S_{gt} is the score of gene g in tissue t; N is the total number of strong enhancers or active promoters associated with the gene g; and P_i is the posterior probability of the associated strong enhancer or active promoter emitted by the ChromHMM model. The RPKM values for each gene quantified using ENCODE's RNA-seq data in 22 tissues were used as a feature for expression data.

The feature for TF binding associated with each gene was calculated by the Makeev lab. They first selected all cistrome regions overlapping -500 bp and +100 bp of TSSs (for each gene, all transcripts from gencode vM15 were considered). Then, the Makeev lab calculated the $-log_{10}$(p-value) of HOCOMOCO motif hits within these cistrome regions (aggregating over all motifs if there were multiple models for a particular TFBS). The respective values for each TF were taken as the TFBS features. The final set of the TFBS features covered all TFs for which we had the ChIP-seq cistrome peaks and a binding motif model (n = 297).

For PPIs, all the protein interactions in mouse were collected from STRING database version 10.5. For a gene g, its PPI connectivity with all strong enhancer and active promoter associated genes in tissue t was calculated as $PPI_{gt} = \sum_{i=1}^{N}(I)$, where N is the total number of enhancer or promoter associated genes in tissue t and I_i is the combined interaction score between gene g and i^{th} gene. Similarly for each gene, its PPI connectivity with all genes known to be associated with the phenotype domain to be predicted was computed as $PPI_{g-phen} = \sum_{i=1}^{M}(I)$, where I_i is the interaction score and M is the total number of known phenotype associated genes from the MGD.

The random forest classifier was implemented in R using randomForest and caret package (Kuhn, 2008). I sought to predict gene-phenotype associations from 13 different phenotypes relevant to at least one tissue type in my dataset. The known gene-phenotype associations from the MGD (top level MP annotations) served as true-positives for the classifier models. The random forest classifier was trained on a subset of genes, where features described above were used as predictor variables and phenotype calls from the MGD as the response variable. This model was used to predict gene-phenotype associations in the remaining set of genes not used in the training of the model. The preProcess function in caret was used to centre and scale all the gene features. Down-sampling was employed on the training data to avoid the impact of class imbalance on

model fitting. Model optimisation across these parameters was performed using k-fold cross validation technique, to choose the model with the best ROC (parameters used: method = 'repeatedcv', number = 5, repeats = 5, metric = 'ROC'). In order to compare the predictive capability of various gene features, 10 different models with different gene feature combinations were built for each phenotype domain (130 models in total). Each of these classifier was assessed by generating ROC and PR curves based on 5-fold cross validation repeated 10 times. For each cross validation run, the true-positive rate (TPR, also called the sensitivity), false-positive rate (FPR) and precision were calculated as follows:

$$TPR = \frac{TP}{TP+FN}$$

$$FPR = \frac{FP}{FP+TN} \tag{4.2}$$

$$precision = \frac{TP}{TP+FP}$$

where TP, FP, TN and FN represents the number of true-positives, false-positives, true-negatives and false-negatives respectively. The cross validation results were then averaged for comparison and reporting purposes. The ROC and PR curves along with their area under the curve were computed using the ROCR and PRROC R packages. The top false positives hits (prediction probability ≥ 0.90) from each model were examined using OpenTarget validation platform to validate the novel predictions from the random forest classifier.

4.4 Discussion

Regulatory elements have been identified in a plethora of cell-types and tissues, however there is limited understanding about their relationship to overall gene function and the resulting disease or phenotype. To gain insights into the mammalian regulatory landscape and its potential impact on phenotypic outcome, I focused the analysis on tissue-specific enhancers. By generating a catalogue of super, typical and weak enhancers in multiple mouse tissues, I systematically investigated their roles in gene function. From multiple aspects such as gene expression, PPI networks and phenotypes, this study provides evidence that SE and TE associated genes share common phenotypic outcomes even though their expression profiles and overall numbers in the genome differ.

A major challenge in the functional characterisation of enhancers is to identify the enhancer target genes. Identifying genes regulated by specific enhancers is not straightforward as an enhancer can often control multiple genes, while in some cases the target gene could be several Mbs away and therefore, not the nearest gene. In this chapter, GREAT was used to associate SEs and TEs to their potential target genes, which do so by computationally defining a 'regulatory domain' for each gene and links a regulatory region to genes within the same domain. GREAT has been widely used in the past to computationally assign TF bound peaks (Benton et al., 2019; Diéguez-Hurtado et al., 2019; Hsu et al., 2019; Nacht et al., 2019) and enhancer associated regions (Hashimoto et al., 2019; Istaces et al., 2019; Li, Kvon, et al., 2019; Li, Yang, et al., 2019; Stone et al., 2019; Zhu et al., 2019) to their potential target genes in various tissues and cell lines. Based on findings in previous chromatin interaction studies, this approach would incorrectly predict or disregard target genes in many cases. Despite of false-positives, analysis based on enrichment or common patterns have shown genome-wide GREAT results to be accurate enough to capture and visualise significant functional patterns associated with the non-coding regions. However, a better alternative to this method would be an integrated approach using Hi-C data to more accurately link enhancers to their target genes. This would also remove noise from the data for subsequent downstream analysis.

SEs have been previously reported to drive high total-expression and regulate tissue-specific expression (Adam et al., 2015; Hnisz et al., 2013; Liu and Lefebvre, 2015; Loven et al., 2013; Siersbæk et al., 2014; Vahedi et al., 2015; Whyte et al., 2013), while TEs have been considered less important regardless of their huge numbers in the genome. To get a more profound understanding of enhancer regulation, the tissue-specific enhancers were classified into SEs (24%) and TEs (76%). Although, SEs constitute approximately a quarter of the total tissue-specific enhancer pool, the SEC

have higher total- and tissue-specific expression compared to other enhancer classes, and the current study extends this association across multiple mouse tissues (Fig. 4.3). Further detailed analysis of tissue-specific expression revealed that while only a fraction (16%) of SEC could be associated with high tissue-specific expression, this was at least 4 times larger than the fraction of genes within TEC (4%) and WEC (3%). However, due to large number of TEs in the genome, TEC contribution towards all levels of tissue-specific expression is substantially more compared to the SEC. This shows that tissue-specific gene regulation is not confined to SEs, and TEs are also involved in tuning the gene expression landscape.

SEs are comprised of dense enhancer clusters spanning large genomic regions and have been shown to be associated with master TFs and other key cell identity genes (Hnisz et al., 2013; Whyte et al., 2013). It was observed that compared to TEs, SEs consists of a large number of constituent enhancers, however, the mechanistic mode of action of these individual constituent enhancers is not well understood. It remains unclear whether the constituent enhancers exert an additive or a more complex cooperative effect on target gene expression. Using these genome-wide enhancer maps, I sought to predict the effect of constituent enhancer density on a gene's total- and tissue-specific expression at a genome-wide scale. This data shows that globally, total- and tissue-specific expression levels are weakly correlated with the number of constituent enhancers. The constituent enhancer density explains a small fraction of the variation in gene expression (total-expression: $r^2 = 0.01$; tissue-specific expression: $r^2 = 0.03$), indicating that there may exist a complex rather than a linear additive relationship between constituent enhancers and target gene expression (Fig. 4.4). Not all constituent enhancers appear to contribute to the transcriptional output with the same strength, suggesting some constituent enhancers may make small contributions therefore helping to fine tune the expression patterns of their associated gene. This prediction is consistent with previous *in vivo* experiments showing deletion of individual constituent enhancers within a SE leads to variable amount of reduction in target gene mRNA levels (Shin et al., 2016; Suzuki et al., 2017). The SE constituents have also been identified to have frequent chromatin interactions amongst themselves (Dowen et al., 2014), suggesting these constituent enhancers may have an effect on one another's contribution towards the target gene transcriptional activity. However, it cannot be precisely extrapolated from this computational analysis that the non-additive relationship between constituent enhancers and gene expression is a result of only cooperative activity between them. Therefore, the possibility that some constituent enhancers may be inactive/false-positives or may have a redundant function in transcriptional activation cannot be ruled out (Moorthy et al., 2017). These redundant enhancers termed as 'shadow enhancers' in *Drosophila*, have been shown to be critical for phenotypic robustness under conditions of environmental or genetic disturbance, and proposed as a mechanism to gain new regulatory functions

during evolution without disturbing the existing robust regulatory networks (Cannavo et al., 2016; Frankel et al., 2010; Hong et al., 2008).

The majority (78%) of SEC was also observed to be associated with one enhancer tissue-type compared to only 27% in TEC, suggesting genes in SEC are likely to be associated with SEs of only one tissue (Fig. 4.5). The number of distinct enhancer tissue-types associated with a gene influences its expression across multiple tissues. Genes associated with a low number of enhancer tissue-types have a tendency to be expressed in a tissue-specific manner, while genes associated with a high number of enhancer tissue-types have a relatively low tissue-specific expression. Altogether, the large number of constituent enhancers and an association with a low number of enhancer tissue-types could possibly explain the up-regulated total- and tissue-specific expression in the SEC.

Prior research has thoroughly investigated the role of SEs in complex traits, showing that disease-causing SNPs are more enriched in SEs of disease-relevant cell-types (Farh et al., 2015; Hnisz et al., 2013; Parker et al., 2013). However, little research has been conducted to systematically examine the effect of SEs and TEs on diseases. In this chapter, I investigated the mammalian phenotype and disease associations of SE and TE associated genes. Both the SEC and TEC were detected to be significantly enriched in phenotypes and diseases in the corresponding tissue-types, emphasising that phenotypes are governed by tissue-specific enhancers. Using phenotyping data from knockout mouse lines of enhancer associated genes, I showed that there is no significant difference in severity and breadth of phenotypes produced from knockouts of SEC and TEC (Fig. 4.11 and 4.12), which highlights the importance of both enhancer classes in disease causation. In addition, no difference in enrichment of mouse essential genes was identified amongst SEC and TEC. Overall, I did not find any significant contrast between the potential phenotypic impact of SEC and TEC, suggesting that functional testing of all enhancers irrespective of categories is fundamental in making any conclusions about their functional significance and phenotypic impact. Although the majority of key cell identity genes and TFs are associated with SEs, the 'peripheral' genes associated with TEs appear to equally contribute towards disease aetiology. A possible explanation to this surprising result is the existence of an 'omnigenic' architecture (Boyle et al., 2017) where regulatory networks are densely inter-related such that TE associated genes expressed in disease-relevant cell-types can collectively impact the regulation of key cell identity genes. To this end, I hypothesised that tissue-specific enhancer associated genes are components of protein complexes involved in aberrant disease-causing biochemical processes and could be potential therapeutic targets. The PPI analysis shows that enhancer associated genes with no prior corresponding tissue-type phenotypic associations preferentially interact with known phenotype-associated genes.

This observation suggests that these enhancer associated genes are potentially involved in the same functional pathway as the known phenotype genes and could serve as novel targets for diseases.

Finally, using a machine learning approach, I systematically evaluated the capability of TSREs and other molecular properties to predict gene-phenotype associations in the mouse (Fig. 4.16). By comparing classifiers trained on different gene features, the classifier with all the gene features combined was found to perform the best in predicting gene-phenotype associations. This analysis also revealed that PPI data have a high predictive capacity to infer mammalian gene-phenotype associations, while regulatory data provides a modest but additive source of information. Further examination of the top scoring false-positive predictions showed their promising application in generating hypothesis about gene function and in identification of potential novel disease targets. Such prediction models can assist in prioritising genes in mouse knockout and genome editing studies. They could also help in selecting the most relevant phenotyping procedures (which often involves costly assays) for transgenic mouse models. It should be noted that the gene features computed from various datasets for the current random forest implementation are not exhaustive. Other genomic and molecular features that could be added to the prediction model includes gene structure information (such as gene length, transcript count, exon count, intron size, UTR length, GC content), splicing information, isoform expression, tissue-specific expression, genetic variation, distance of regulatory element from gene, and protein expression. It would be interesting to see whether incorporating these gene features could improve the classifier's performance. Another limiting factor contributing to the poor performance in predicting some phenotype domains (such as adipose tissue phenotype) is the low number of known phenotype associations to serve as true positives in the classifier training phase. To predict genes associated with such phenotypes appears to be relatively difficult with the current number of annotations, but the accuracy of such models would improve as the number of known annotations increases in future.

In summary, the findings in a diverse range of mouse tissues from this chapter present opportunities for molecular experiments to investigate regulatory mechanisms in mouse models of human diseases. Further *in vivo* studies are required to more thoroughly understand the impact of enhancers on gene expression. Ideally, one would disrupt all the enhancer loci individually and examine its effect on their target gene expression. This would be extremely time consuming given the large number of enhancers in the mouse genome, but with the swift advancement in genome editing techniques, a large proportion of these potential enhancers might be characterised in the near future.

Chapter 5

Assessing the role of *Zfhx3* as a circadian regulator in the SCN

In this chapter, I describe the RNA-seq and ChIP-seq analysis of a novel mouse model Short Circuit (*Sci*) with a circadian phenotype caused by a mutation in the *Zfhx3* gene. During this project, I worked closely with the neurobehavioral genetics group (Nolan lab) at the MRCHI. The transcriptome analysis of *Sci* described here has been published in the following article:

Parsons, M. J., M. Brancaccio, **S. Sethi**, E. S. Maywood, et al. (2015). "The Regulatory Factor ZFHX3 Modifies Circadian Function in SCN via an AT Motif-Driven Axis". In: *Cell* 162.3, pp. 607-621. ISSN: 0092-8674. DOI: 10.1016/j.cell.2015.06.060.

5.1 Introduction

The neurobehavioral genetics group at the MRCHI focuses on understanding the genetics of mammalian behaviour and circadian rhythms, where they use mouse models to investigate genes that regulate our inner body clock. Circadian (from the latin words *circa* and *diem*, meaning 'about a day') rhythms are 24 hour cycles of internal molecular clocks that influence mammalian physiology and behaviour. Many important functions at the molecular and behavioural level can be influenced by circadian rhythms such as food intake, metabolism, body temperature, DNA repair mechanisms and sleep cycles. Abnormal circadian rhythms have been associated with many disorders like obesity, neurodegenerative and mood disorders, depression, insomnia, and high risk of breast cancer and metabolic diseases (Bass and Lazar, 2016). These rhythms are controlled by molecular clocks which are present in most cells, further regulated by a 'master clock' located in the suprachiasmatic nucleus (SCN) (Ralph et al., 1990).

The master clock in the SCN receives light signals from retina and synchronises the molecular clocks to the external environment and to each other. This process involves an interconnected network of TFs, genes and regulatory motifs such as the E-box, D-box and *Rev-erbα*/ROR-binding elements (RREs) (Ukai and Ueda, 2010).

The core circadian gene network in mammals consists of two transcriptional-translational feedback loops (TTFLs) (reviewed in Partch et al. (2014)) (Fig. 5.1). The TTFL comprises of four important clock proteins: *Clock* and *Bmal1* acting as activators; and *Per* and *Cry* acting as repressors. *Clock* and *Bmal1* activate the transcription of *Per* and *Cry* genes (along with other clock genes) via the E-box motif. *Per* and *Cry* protein products in turn translocate to the nucleus and interact with *Clock:Bmal1* complex to negatively regulate them, hence restraining their own further activation. As *Per* and *Cry* proteins get degraded, the restriction on *Clock:Bmal1* is inactivated and the cycle starts again. The casein kinases (*CKI*) determines the rate at which *Per* and *Cry* are degraded and enter the nucleus. This positive and negative loop is connected to another TTFL comprising of a clock gene called *Rev-erbα*. Similar to *Per* and *Cry* genes, *Rev-erbα* is also activated by *Clock* and *Bmal1* complex via the E-box motif in its promoter. Interestingly, *Rev-erbα* protein in turn acts as a transcriptional repressor for *Bmal1* by binding retinoic acid-related orphan receptor response elements (ROREs) in *Bmal1* promoters. Hence, due to *Bmal1* involvement in activation of *Rev-erbα*, *Bmal1* indirectly plays a role in repressing its own transcription. The presence of such interlocking TTFLs creates a stable circadian model against environmental changes and also provides a mechanism to create phase delays in the activation of circadian genes to suit the gene expression requirement according to the local physiology. These TTFLs organise the molecular clocks in individual cells, but within the SCN, the cells are synchronised through intercellular coupling regulated by neuropeptides such as vasoactive intestinal peptide (*Vip*) (Aton et al., 2005) and gastrin-releasing peptide (*Grp*) (Brown et al., 2005). Furthermore, other peptides like neuromedin S (*Nms*) and prokineticin 2 (*Prok2*) have been identified to be important for circadian signalling between the SCN and other parts of the brain (Lee et al., 2015; Prosser et al., 2007).

In the recent years, circadian rhythms research has benefited from various gene discovery approaches, including ENU mutagenesis screens to identify novel genes involved in the regulation of mammalian circadian circuit. This approach involves treating the mice with ENU, a chemical mutagen which causes random point mutations in the genome, and then screening the offspring for phenotypes (Nolan et al., 2000). From the ENU mutagenesis screen of circadian phenotypes at the MRCHI, the Nolan lab identified a mouse model in the G1 animals displaying a dominant shortened circadian period ($T_{DD} = 23.0 \pm 0.05$ hr; T_{DD}: behavioural circadian period of the animals in constant darkness) as compared to the population mean ($T_{DD} = 23.6 \pm 0.08$ hr) (Fig.

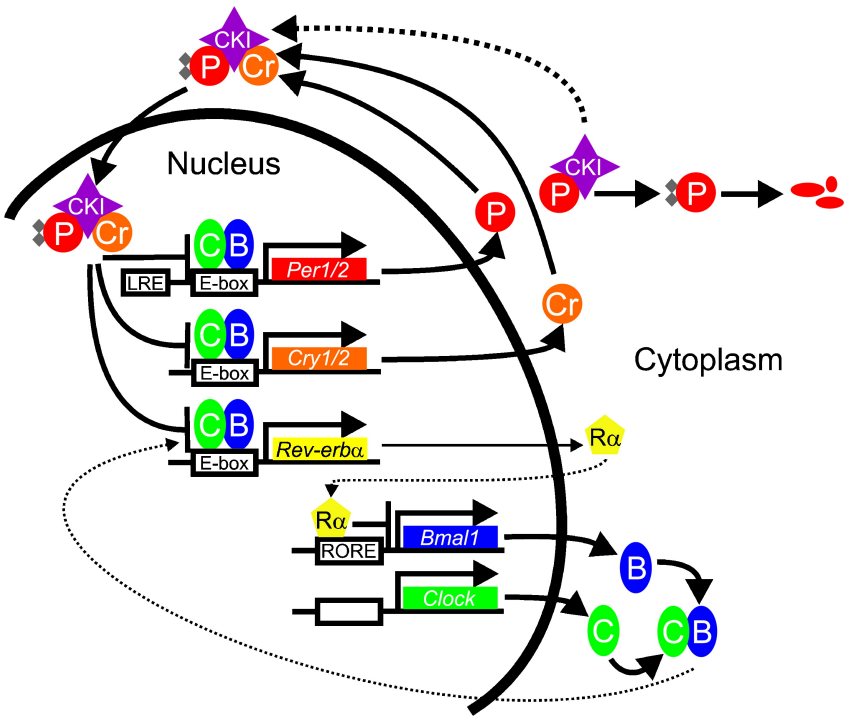


Fig. 5.1 A schematic illustration of the mammalian core circadian clock. A diagram displaying the transcriptional-translational feedback loops (TTFLs) in the mammalian circadian gene network. P: *Per*, Cr: *Cry*, B: *Bmal1*, C: *Clock*, Rα: *Rev-erbα*, LRE: light-responsive elements in the *Per* promoters. Grey diamonds display phosphorylation. Figure taken from Lowrey and Takahashi (2004).

5.2A-B). Using positional candidate analysis, the observed phenotype was associated to the genomic location between 107.67 Mb and 110.57 Mb on mouse chromosome 8 comprising of 25 genes. The coding regions of all these genes were scanned for mutations using Sanger sequencing, which detected a point mutation in zinc-finger homeobox 3 (*Zfhx3*) gene. The mutation was identified in exon 9 of *Zfhx3*, substituting a phenylalanine to valine in a highly conserved region. This mutation was named as Short Circuit (*Sci*).

Zfhx3 (also known as *Atbf1* because of its property to bind AT rich motif) is a large TF comprising of multiple zinc finger and homeodomains, known to be involved in multiple biological functions (Yasuda et al., 1994). *Zfhx3* controls neuronal and myogenic differentiation, acts as a tumour suppressor in some cancers, and its knockout in mice leads to developmental defects (Berry et al., 2001; Jung et al., 2005; Kaspar et al., 1999; Sun et al., 2012). Moreover, *Zfhx3* expression is found to be highly enriched in the adult SCN (Lein et al., 2007). The homozygous *Sci* mutation caused lethality during embryonic development, hence only heterozygous adult animals ($Zfhx3^{Sci/+}$) could be further investigated. The Nolan lab performed qPCR to examine the mRNA expression of the core circadian genes across the light-dark cycle in the SCN, which revealed no significant differences between the $Zfhx3^{Sci/+}$ and $Zfhx3^{+/+}$ animals (Fig. 5.2C). This

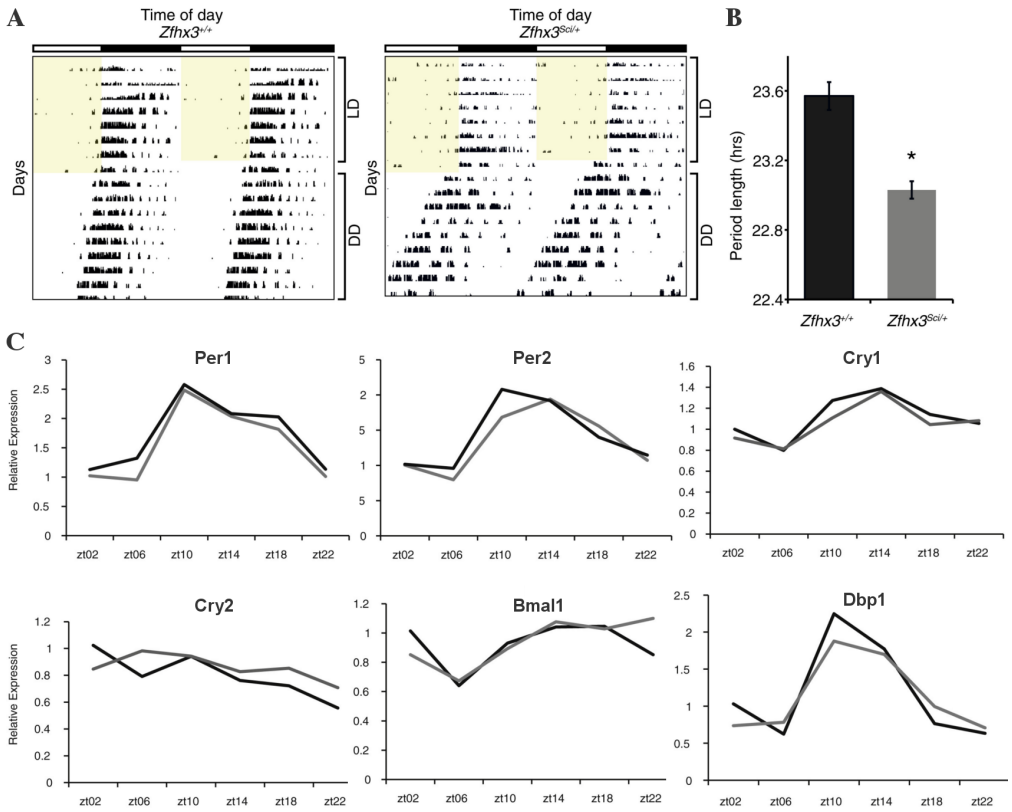


Fig. 5.2 Overview of the short circuit (*Sci*) phenotype. (A) Actograms displaying the wheel running activity in *Zfhx3*$^{Sci/+}$ and *Zfhx3*$^{+/+}$ mice. The mice were kept on a light-dark (LD) cycle for 7 days, followed by 2 weeks in constant darkness (DD). Yellow shaded area shows the duration when lights were on. (B) Bar plot displaying the free running period length of *Zfhx3*$^{Sci/+}$ and *Zfhx3*$^{+/+}$ mice in constant darkness (n = 6) (* denotes p = 0.0009). (C) mRNA expression of the core circadian genes in the SCN of *Zfhx3*$^{Sci/+}$ (grey lines) and *Zfhx3*$^{+/+}$ (black lines) at six time points (n = 4, p > 0.2, ANOVA). Figure taken from Parsons et al. (2015).

suggested that the *Sci* phenotype is potentially caused by an altered *Zfhx3*-dependent pathway rather than a TTFL-dependent effect. To understand how the *Sci* mutation in the *Zfhx3* disrupts the SCN circadian period, the transcriptome was sequenced to analyse the gene expression levels between the *Zfhx3*$^{Sci/+}$ and *Zfhx3*$^{+/+}$ mice at two circadian time points: Zeitgeber Time (ZT) 3 and 15 (ZT is a standard 24 hour notation of the circadian cycle in which ZT=0 represents start of the day or the light phase, and ZT=12 represents start of the night or the dark phase). The Nolan lab further performed a ChIP-seq of the TF *Zfhx3* at ZT3 and ZT15, to study its genome-wide binding patterns and determine its associated functional pathways in the SCN. In this study, I describe the transcriptome analysis of the mice exhibiting abnormal circadian cycles to identify the impact of the *Zfhx3*^{Sci} mutation on the transcriptional targets of *Zfhx3*. I further investigate the regulatory networks and the mechanism potentially contributing to the *Sci* phenotype by predicting molecular complexes in the *Zfhx3* transcriptional network. Finally, I map the *in vivo* binding of *Zfhx3* in the SCN to inspect its binding affinity between two circadian time points and identify its DNA binding motif. The results from

this study provide evidence of a novel role of *Zfhx3* in maintaining circadian oscillations via a network of neuropeptides in the SCN.

5.2 Results

5.2.1 Effect of *Zfhx3*^{Sci} mutation on gene expression

In order to examine the expression changes in *Zfhx3*$^{Sci/+}$ animals and identify the transcriptional targets of *Zfhx3*, RNA from the SCN tissue of mutant (*Zfhx3*$^{Sci/+}$) and wild type (*Zfhx3*$^{+/+}$) mice was sequenced at ZT3 and ZT15 (n = 3). For RNA-seq analysis, I developed a pipeline integrating multiple statistical methods to identify differential expression between wild type and mutant mice. In brief, the pipeline aligns RNA-seq reads to the genome and identifies differentially expressed genes using three methods namely EdgeR (Robinson et al., 2010), DESeq (Anders and Huber, 2010) and Cufflinks (Trapnell et al., 2012). A further filtering criteria was employed ($\log_2$ fold change $> \pm 1$, $q < 0.05$ in at least two methods) to capture highly significant gene expression changes. This approach detected 28 genes to be differentially expressed at least at one time point (Fig. 5.3A). Of these 28 genes, 19 were differentially expressed at ZT3 and 13 at ZT15 (Fig. 5.3B), while 4 genes were altered at both time points. Most of the differentially expressed genes (17/28) were down-regulated in the *Zfhx3*$^{Sci/+}$ mice. Interestingly, the down-regulated genes included a number of neuropeptides and their receptors (such as *Vip, Vipr2, Avp* and *Prokr2*) previously known to be associated with circadian rhythms (Reghunandanan and Reghunandanan, 2006) (Fig. 5.3C). However, no significant difference in the expression of core clock genes (such as *Per1, Per2, Cry1* and *Cry2*) was observed.

Previous work identified that *Zfhx3* regulates genes via an AT rich motif (Yasuda et al., 1994). Therefore, I examined whether genes affected by the *Zfhx3*$^{Sci/+}$ mutation contains the AT motif sequence in their promoter region. Other well known circadian related motifs such as the E-box, D-box and RRE were also included in this analysis (Fig. 5.4). The core promoter regions (450 bp upstream and 50 bp downstream of TSSs) of all differentially expressed genes ($q < 0.05$ in at least one analysis method, n = 168) were searched for the presence of these motif sequences using Pscan (Zambelli et al., 2009). This approach identified 39% (66/168) of the differentially expressed genes to contain a predicted AT motif (Pscan score > 0.88), while the circadian motifs E-box, D-box and RRE were absent (Fig. 5.4).

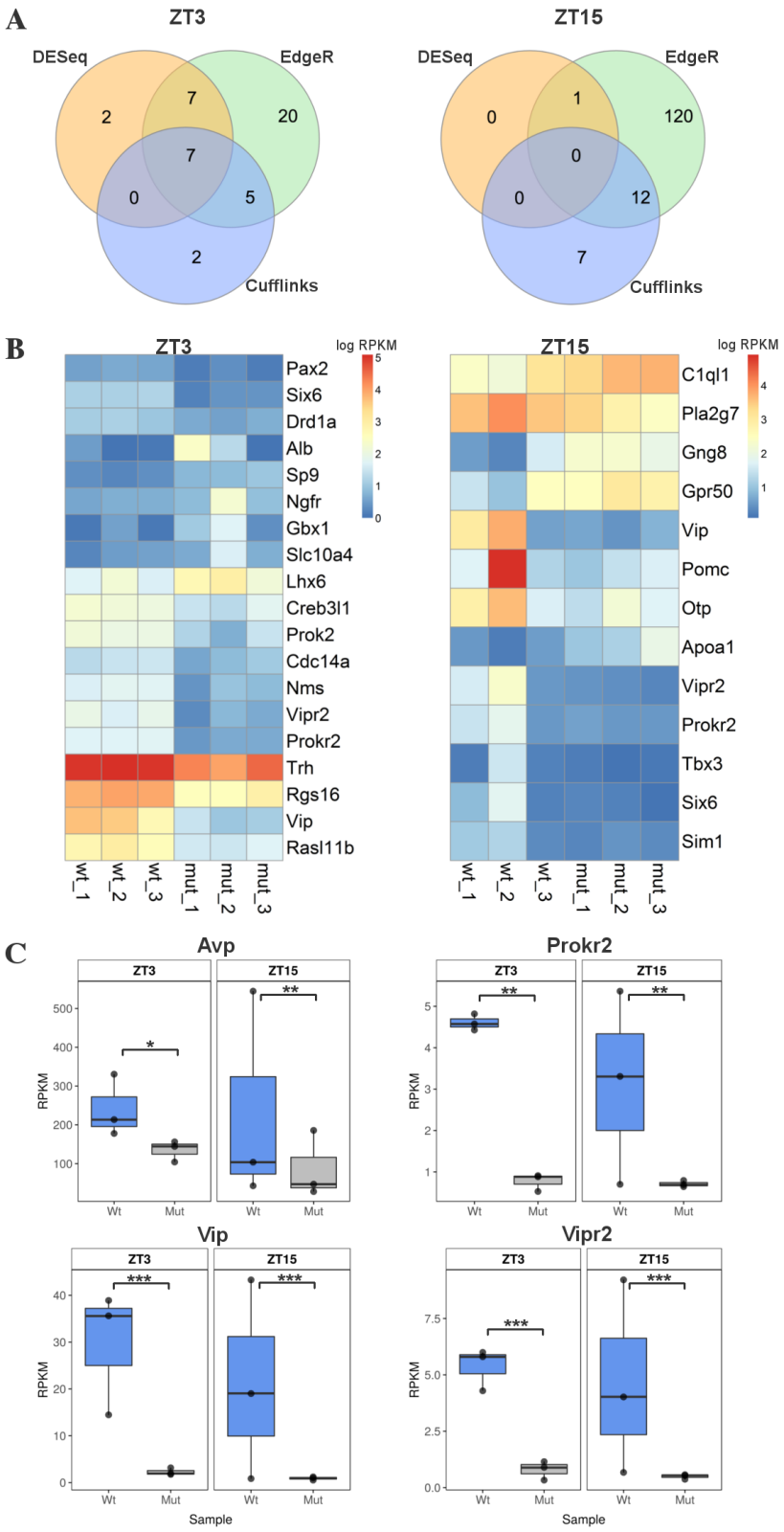


Fig. 5.3 SCN genes differentially expressed between *Zfhx3*$^{Sci/+}$ and *Zfhx3*$^{+/+}$. (A) Venn diagrams displaying the number of differentially expressed genes ($q < 0.05$, $\log_2$ fold change $> \pm 1$) detected by each method. (B) Heatmaps showing the expression change in genes identified to be differentially expressed by at least two analysis methods. The columns in the heatmap show gene expression in wild type (wt) and mutant mice (mut) across different replicates. (C) Box plots displaying decreased expression of neuropeptides in the mutant mice ('***' denotes $q < 0.001$, '**' denotes $q < 0.01$, '*' denotes $q < 0.05$; wt: *Zfhx3*$^{+/+}$; mut: *Zfhx3*$^{Sci/+}$).

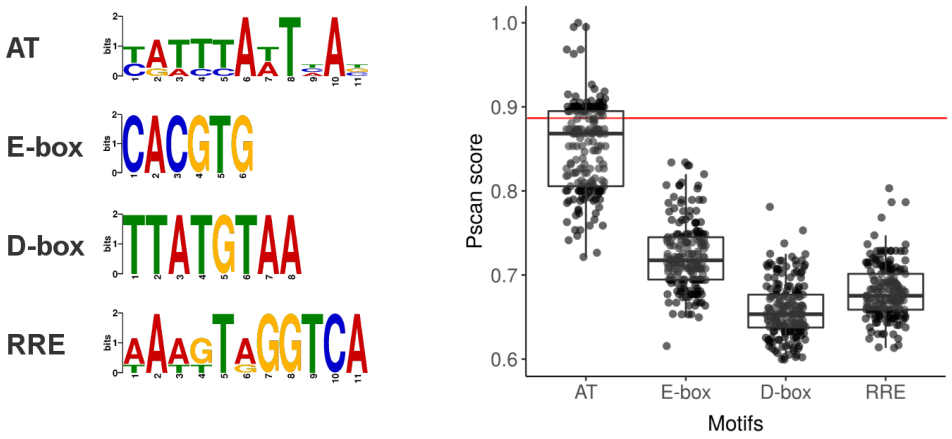


Fig. 5.4 Motif enrichment in differentially expressed genes associated with *Zfhx3*$^{Sci/+}$ mutation. Left panel: positional weight matrix (PWM) logos of motifs used in the analysis. Right panel: box plot displaying the Pscan scores of each motif for the promoter regions of differentially expressed genes. Pscan calculates a score for each predicted motif site which represents the sequence similarity between the motif sequence and the genomic sequence at that site. Red line represents the threshold for Pscan score used here.

5.2.2 Dissecting functionally distinct modules in *Zfhx3*$^{Sci/+}$ network

In order to investigate the functional pathways affected in *Zfhx3*$^{Sci/+}$ mice, I examined the protein-protein interactions (PPIs) amongst differentially expressed genes ($q < 0.05$ in at least one analysis method, $n = 168$) associated with the *Zfhx3*$^{Sci/+}$ mutation using STRING. This revealed 69 (out of 168) genes to have at least one PPI (Fig. 5.5). To identify any potential functional modules in this protein interaction map, I applied a graphical clustering method called MCODE (Molecular Complex Detection) (Bader and Hogue, 2003), which identifies densely linked regions in a network by clustering the nodes on the basis of their interconnectivity. These clusters could correspond to functional complexes in a network. MCODE characterised the network into four modules, with module 1 (comprising of 21 genes) attaining the highest connectivity score of 10.1, followed by module 2 (comprising of 10 genes) with a score of 5.5, and module 3 (comprising of 6 genes) and 4 (comprising of 10 genes) with a score of < 2. GO enrichment analysis revealed these modules to be involved in distinct biological functions; module 1 comprised of genes associated with neuropeptide signalling activity (corrected $p = 1.4 \times 10^{-13}$) (Fig. 5.6A), module 2 comprised of genes related to ribosome function (corrected $p = 8.6 \times 10^{-7}$) (Fig. 5.6B), module 4 contained genes associated with cholesterol homeostasis (corrected $p = 5.7 \times 10^{-3}$) (Appendix B.6), whereas no GO terms were identified to be significantly enriched in module 3.

Interestingly, the majority of genes (15/21) in module 1 had decreased expression in *Zfhx3*$^{Sci/+}$ mice, including that of neuropeptides and their receptors like *Avp, Vip, Vipr2, Prok2, Prokr2, Grp* and *Nms*. The neuropeptides *Vip* and *Grp* have been previously

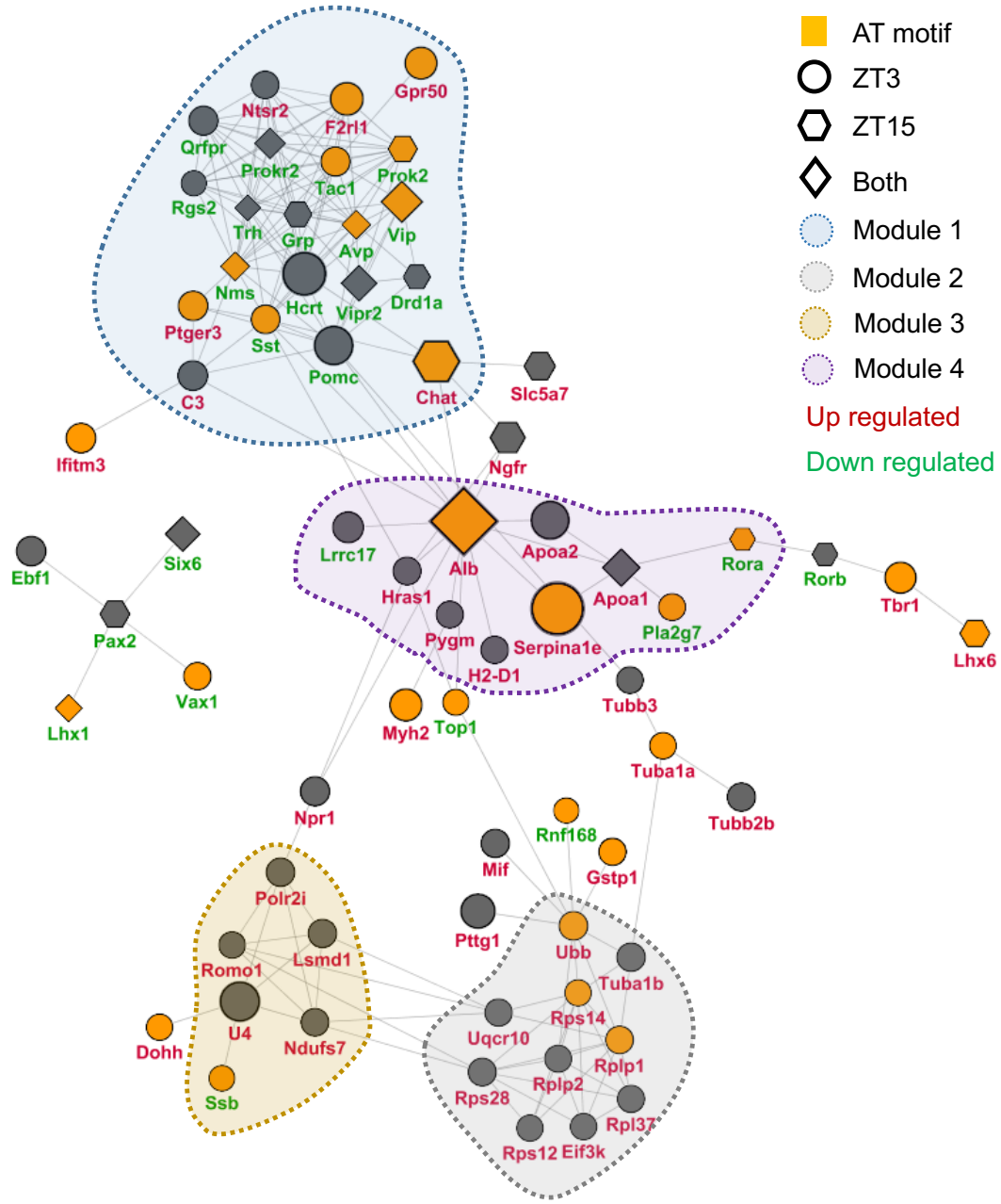


Fig. 5.5 PPI map of differentially expressed genes in *Zfhx3*$^{Sci/+}$ mice. A PPI network amongst genes detected to be differentially expressed in *Zfhx3*$^{Sci/+}$ compared to *Zfhx3*$^{+/+}$. The nodes in the network represent genes while edges represent PPIs. The shape of the node depicts the time point at which the gene was identified to be differentially expressed, while the node size represents the corresponding gene expression fold change. Genes predicted to have an AT motif sequence in their promoters are coloured in orange. The clusters represent functional modules in the network identified by MCODE.

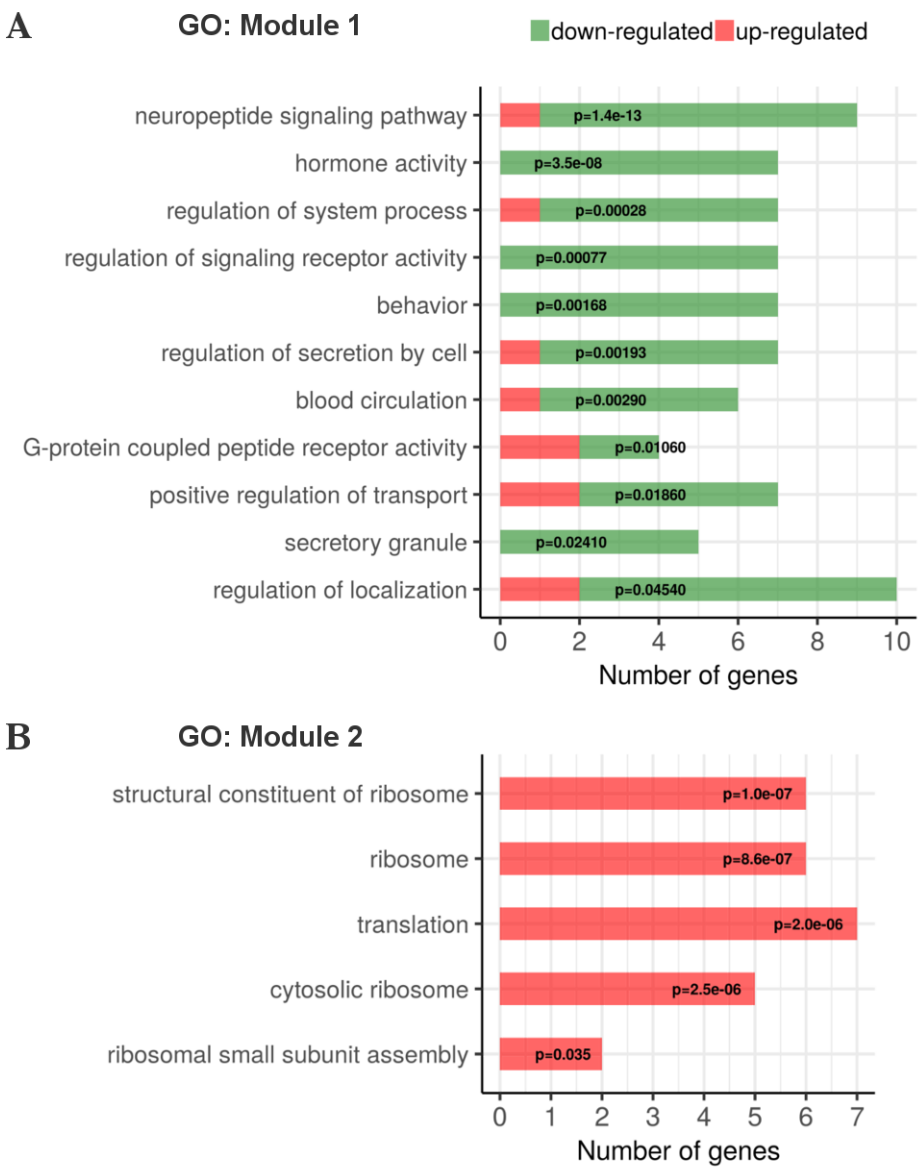


Fig. 5.6 GO enrichment analysis of functional modules in the *Zfhx3*$^{Sci/+}$ network. Bar plots displaying the over-represented GO terms amongst genes in (A) module 1 and (B) module 2. Down-regulated and up-regulated represents genes with decreased and increased expression in *Zfhx3*$^{Sci/+}$ mice respectively.

demonstrated to be involved in the regulation of firing of the SCN neurons (Aton et al., 2005; Brown et al., 2005). The decrease in the mRNA expression of *Vip* and *Grp* in the *Zfhx3*$^{Sci/+}$ SCN was experimentally confirmed by qPCR (Appendix A.11A). Furthermore, 6 out of the 15 down-regulated genes in module 1 had a predicted AT motif binding site in their promoter regions (Pscan score > 0.88, Fig. 5.5), while the circadian related E-box, D-box and RRE motifs were absent. These genes included the down-regulated neuropeptides *Vip, Avp* and *Prok2*. To experimentally validate the binding of *Zfhx3* to the promoters of *Avp* and *Vip*, the Nolan lab performed a quantitative ChIP in the SCN tissue from *Zfhx3*$^{+/+}$ mice. Both *Avp* and *Vip* showed significantly

higher immunoprecipitated DNA compared to the *Gapdh* control promoter region (Appendix A.11B). These results suggest that *Zfhx3* directly regulates the circadian related neuropeptide network via the AT motif. Contrary to module 1, all the genes in module 2 had significantly increased expression in $Zfhx3^{Sci/+}$ compared to $Zfhx3^{+/+}$ (Fig. 5.5). The majority of genes in module 2 are linked to ribosomal function (Fig. 5.6B) indicating that *Zfhx3* may also be involved in the regulation of ribosomal proteins.

Overall, the $Zfhx3^{Sci/+}$ mutation appears to cause a decrease in the ability of *Zfhx3* to activate transcription of circadian related neuropeptides via the AT motif, leading to the *Sci* phenotype. To test this hypothesis, the Nolan lab cloned the AT motif into the pGL3-Enhancer Luciferase Reporter Vector and co-transfected it with an expression vector containing recombinant *Zfhx3*, with or without the *Sci* mutation in HEK293 cells. The $Zfhx3^{+}$ was observed to exhibit higher activation compared to the empty vector, whereas $Zfhx3^{Sci}$ activation was equal to the empty vector (Appendix A.11C). This suggests that $Zfhx3^{Sci}$ has a diminished ability to activate transcription via the AT motif. Furthermore, the Nolan lab also cloned module 1 gene promoters into the pGL3-Enhancer Luciferase Reporter Vector. Three types of genes were selected from module 1 for this experiment: (1) with strong predicted AT motif (*Avp*, *Vip*; Pscan score >0.88); (2) with moderate predicted AT motif (*Grp*, *Prokr2*; Pscan score >0.80); and (3) without the AT motif (*Drd1a*, *Vipr2*; Pscan score <0.80). These reporters were then co-transfected with $Zfhx3^{Sci}$ or $Zfhx3^{+}$ expression vectors (Appendix A.11D). Reporters containing the strong predicted AT motif exhibited increased activation compared to the empty vector for both $Zfhx3^{Sci}$ or $Zfhx3^{+}$ over-expression, but activation by $Zfhx3^{+}$ was significantly higher than that of $Zfhx3^{Sci}$. For reporters containing the moderate predicted AT motif, $Zfhx3^{+}$ showed increased activation of *Prokr2*, but not with *Grp*, while reporters without the predicted AT motif did not show any notable activation. Moreover, to investigate whether *Zfhx3* driven transcription is dependent on the AT motif, the Nolan lab mutated three conserved residues (residues 6, 8, and 10) within *Avp* and *Vip* promoter constructs. The reporters with mutated AT motif showed significantly lower ability of transcriptional activation (Appendix A.11E). These findings further convey the important role of the AT motif in $Zfhx3^{+}$ activation.

5.2.3 Investigating the *Zfhx3* regulome in the SCN

TFs function by binding to DNA within regulatory regions such as enhancers and invoking mechanisms to activate or repress their target gene expression. Identification of genome-wide localisation of TFs can help to unveil the molecular mechanisms underlying their regulatory function and identify specific TF interactions. Therefore, in order to further study the regulatory network of *Zfhx3* in the SCN, ChIP-seq was used to identify *in vivo* genome-wide binding of *Zfhx3*. The ChIP-seq experiment was performed using the SCN from $Zfhx3^{+/+}$ mice at ZT3 and ZT15 time points ($n = 1$). The ChIP-seq analysis (see methods 5.3.4) identified 27,438 and 27,178 significant peaks at ZT3 and ZT15 respectively ($q < 0.01$). Genomic view of few of the *Zfhx3* ChIP-seq peaks identified in the SCN are shown in Fig. 5.7. The average ChIP-seq signal across all the significant peaks shows that *Zfhx3* have similar ChIP-seq intensity and binding profile at ZT3 and ZT15 (Fig. 5.8A-B). Comparing the *Zfhx3* peaks with genomic features revealed that the majority ($\sim$61%) of *Zfhx3* binding in the SCN occurs at promoters (within ≤ 1 kb of TSSs), followed by $\sim$15% in distal intergenic regions and $\sim$9% within 1-3 kb upstream of TSSs (Fig. 5.8C). This shows that *Zfhx3* preferentially binds to promoter regions. Next, I examined what fraction of the *Zfhx3* binding in the SCN occurs within active promoters and strong enhancer annotations (identified using ChromHMM in chapter three) in different brain regions; cerebellum, cortex and whole brain. Overall, $\sim$59% of the total genome covered by *Zfhx3* peaks in the SCN overlaps with active promoters in the cerebellum, cortex or whole brain (permutation test, $p < 10^{-3}$). Whereas only 2.4%, 3.6% and 4.35% overlaps with strong enhancers in the cerebellum, cortex and whole brain respectively. A detailed comparison between the number of *Zfhx3* peaks, and active promoters and strong enhancers in different tissues is shown in Fig. 5.8D-E.

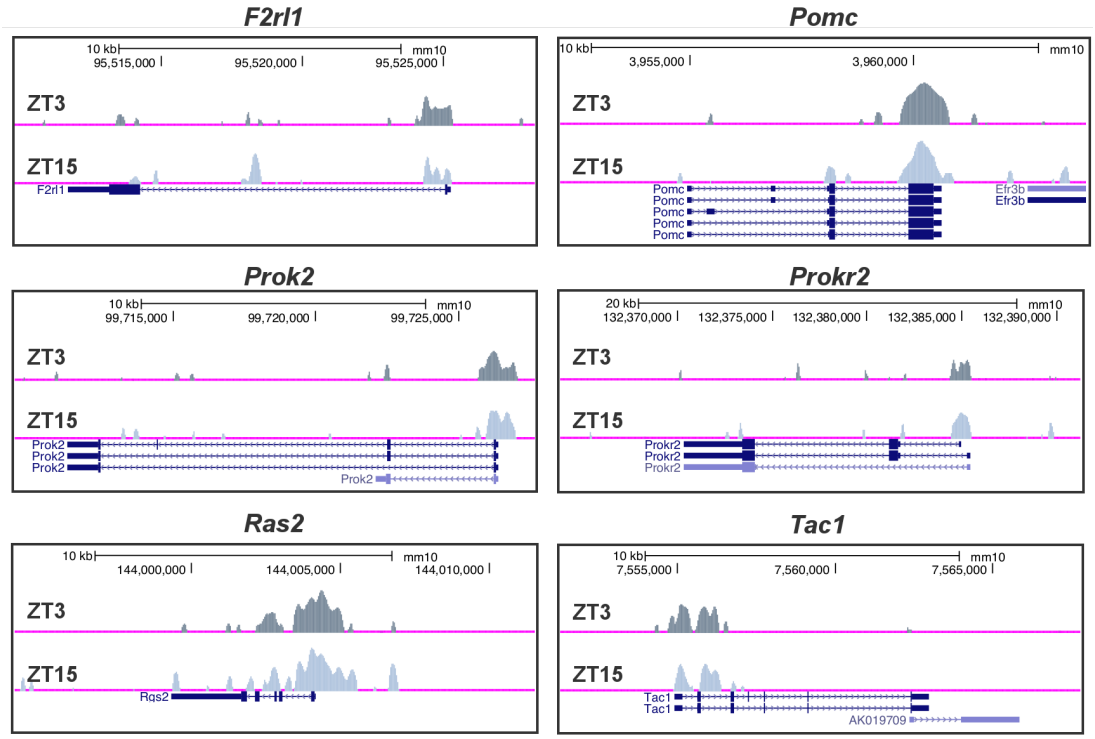


Fig. 5.7 Genomic view of *Zfhx3* ChIP-seq peaks in the SCN. Genome browser snapshots of *Zfhx3* ChIP-seq binding profile. The tracks displayed are input normalised ChIP-seq signal of *Zfhx3* peaks in the SCN at ZT3 and ZT15.

Next, I associated the top 1,000 *Zfhx3* bound ChIP-seq peaks at ZT3 and ZT15 with their potential gene targets using GREAT (McLean et al., 2010). This resulted in 977 and 986 potential *Zfhx3* gene targets at ZT3 and ZT15 respectively. In order to examine the functional roles of these genes in the SCN, GO enrichment analysis was performed. For this purpose, the target genes identified at both time points were combined resulting in 1,216 unique genes. The GO enrichment analysis show that these genes are involved in biological processes such as cellular metabolic process (corrected $p = 10^{-53}$), intracellular receptor signalling pathway (corrected $p = 10^{-2}$), neuron fate commitment (corrected $p = 10^{-2}$); and molecular functions such as nucleic acid binding (corrected $p = 10^{-55}$), transcription regulator activity (corrected $p = 10^{-16}$) and structural constituent of ribosome (corrected $p = 10^{-2}$) (Fig. 5.9). Interestingly, GO terms enriched amongst these genes such as receptor signalling pathway and structural constituent of ribosome, were also enriched amongst differentially expressed genes associated with the $Zfhx3^{Sci}$ mutation.

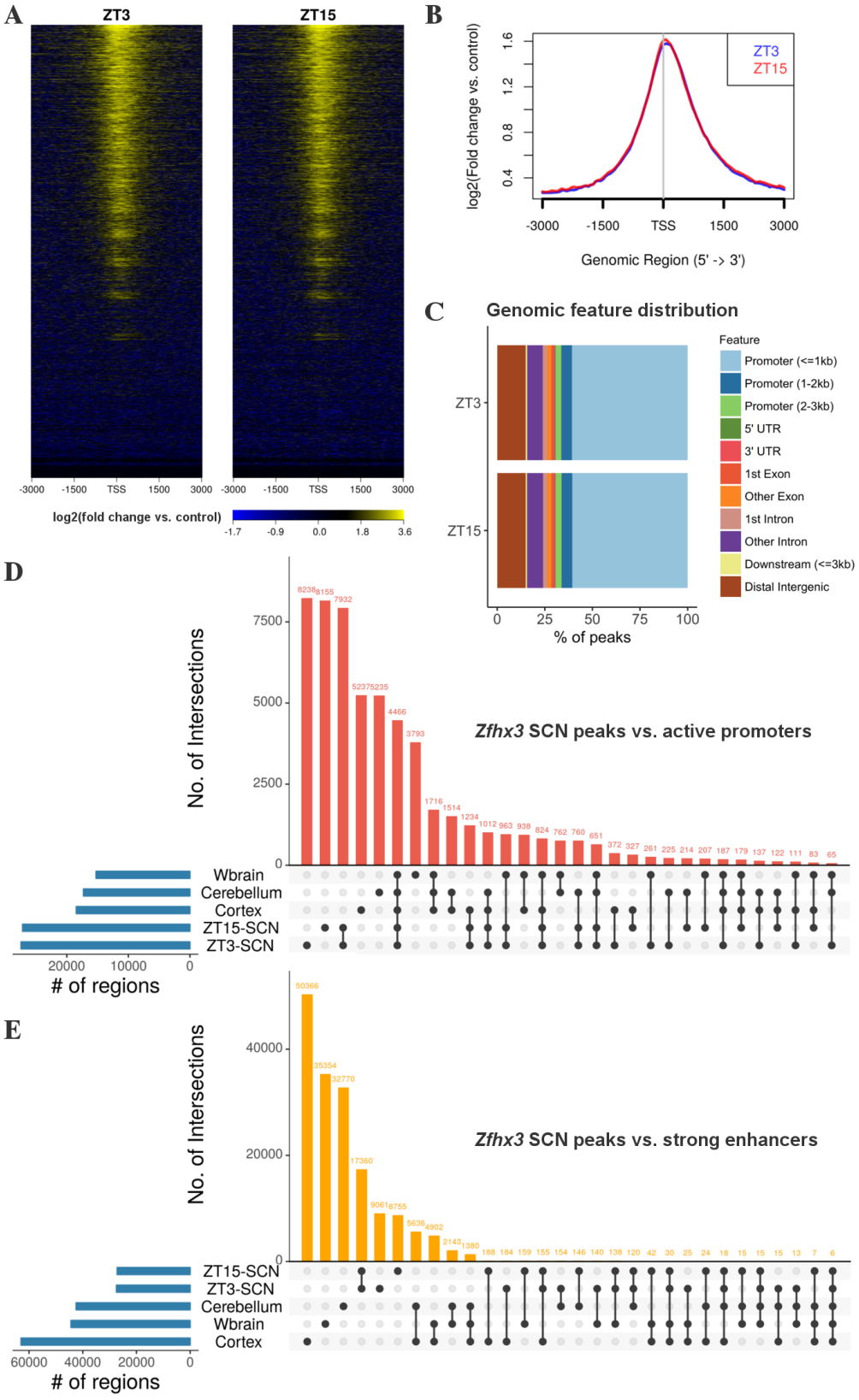


Fig. 5.8 Overview of *Zfhx3* binding profile in the SCN. (A) Heatmaps showing the binding intensity of *Zfhx3* over TSSs and the surrounding 3 kb region. (B) Comparison of average ChIP-seq density profile of *Zfhx3* between ZT3 and ZT15. (C) Distribution of *Zfhx3* binding in the SCN with respect to known genomic features. (D-E) Upset plots displaying the intersection of *Zfhx3* peaks in the SCN with active promoter and strong enhancer regions in the cerebellum, cortex and whole brain.

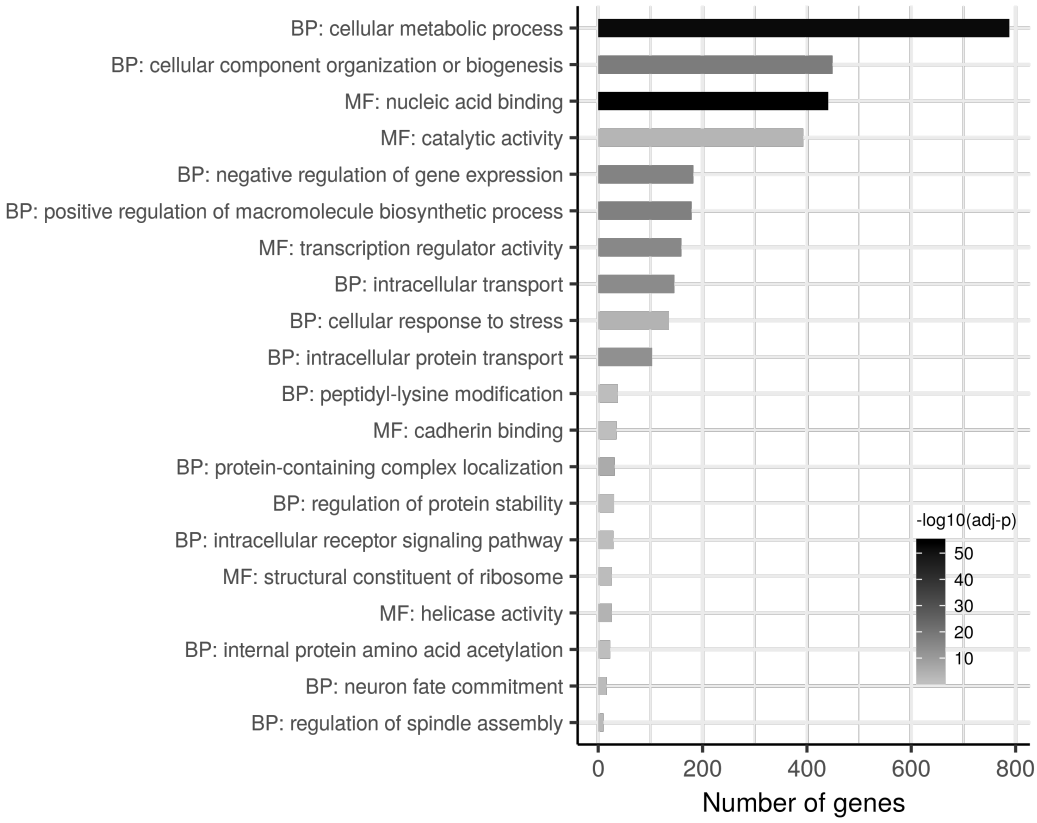


Fig. 5.9 GO enrichment analysis of genes associated with *Zfhx3* ChIP-seq peaks. Bar plots displaying the significantly over-represented GO terms amongst genes associated with the top 1,000 *Zfhx3* ChIP-seq peaks at ZT3 and ZT15. The gradient colour of the bars show the log transformed corrected p-values associated with the GO terms. BP: biological processes; MF: molecular function.

5.2.4 Differential *Zfhx3* binding between ZT3 and ZT15

TF binding patterns and their gene regulatory networks are dynamic. Interactions between various TFs and their gene targets along with external signals causes the gene expression landscape to change with respect to time (Swift and Coruzzi, 2017). Due to this reason, it is important to study the binding and function of TFs across different time points, specially in the SCN where many circadian regulators oscillate in 24 hour cycles. Therefore, to identify any significant transient events in *Zfhx3* binding, I analysed its binding patterns between ZT3 and ZT15. Comparing the genomic coordinates of *Zfhx3* binding peaks between ZT3 and ZT15 revealed that ∼85% (23,453) of the peaks (at $q < 0.01$) are common (with at least 1 bp overlap), with ∼74% (20,319) of the ZT3 peaks covering at least 50% of the genomic area within ZT15 peaks. In order to identify *Zfhx3* binding sites with significantly different binding affinity between ZT3 and ZT15, I compared the ChIP-seq signal at their respective peak regions. However, this analysis identified no significant differentially bound regions between ZT3 and ZT15 ($p < 0.05$, $\log_2$ fold change $> \pm 1$) (Fig. 5.10). Despite no *Zfhx3* peaks attaining a statistically

significant p-value ($p < 0.05$), 82 peaks (associated with 74 genes) were detected to be differentially bound with a $\log_2$ fold change $> \pm 1$. Using GO enrichment analysis, I examined whether these 74 genes associated with differentially bound peaks have common biological processes in the SCN. However, no GO terms were significantly enriched amongst these genes ($q < 0.05$). These results indicate that *Zfhx3* binding affinity is similar between ZT3 and ZT15 in the SCN.

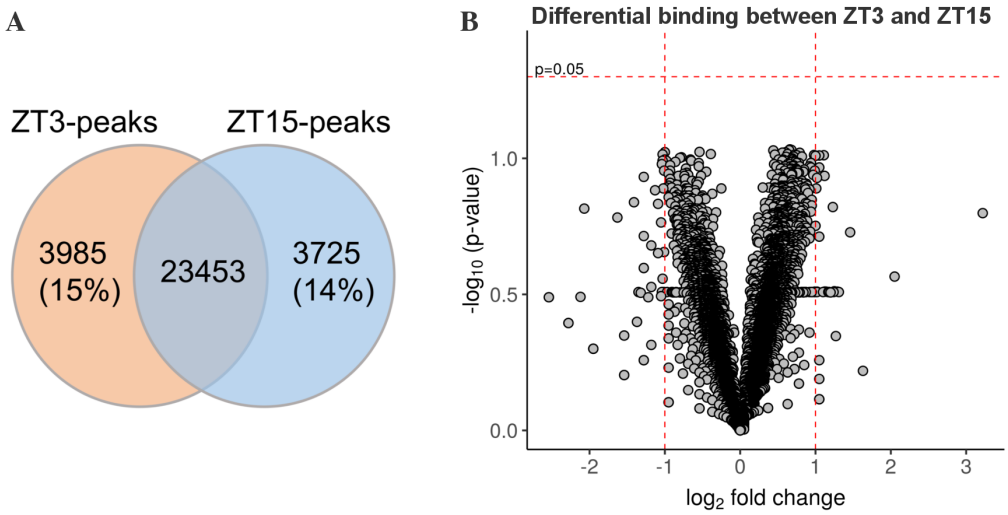


Fig. 5.10 Differential binding analysis of *Zfhx3* activity between ZT3 and ZT15. (A) Venn diagram displaying the overlap between the genomic coordinates of *Zfhx3* peaks (at $q < 0.01$) at ZT3 and ZT15. (B) Volcano plot displaying the p-values and fold changes obtained from differential binding analysis of *Zfhx3* peaks between ZT3 and ZT15.

5.2.5 Identifying *Zfhx3* binding motif in the SCN

TFs comprise of one or more DNA binding domains which recognise a small set of specific DNA sequences (called a motif) and then bind to it. Motifs can aid in the study of protein structure of the TF, and in the identification of co-expressed genes and co-regulators involved in the same functional pathway. Most importantly, since the majority of the functional binding sites of a TF should contain its consensus motif, the motifs can help to accurately predict the true-positive binding peaks in a ChIP-seq experiment (Maurano et al., 2012). To identify the DNA motif sequence recognised and bound by *Zfhx3* protein in the SCN, *Zfhx3* binding peaks were analysed to detect over-represented motif sequences within it. For this analysis, the top 1,000 peaks (based on q-value) from each time point were used. Additionally, I sought to examine whether *Zfhx3* motif binding patterns are similar in promoter and enhancer regions. In order to do this, *Zfhx3* peaks were classified into two groups; peaks binding within 2 kb upstream and 200 bp downstream of a known TSS were considered to be occurring in promoter regions, whereas the remaining peaks were considered to be within enhancer regions.

This resulted in ∼64% of the peaks to be classified within promoter regions at both time points. The top 1,000 peaks (based on q-value) within promoter and enhancer regions were then analysed for motif enrichment (see methods 5.3.5).

The motif analysis identified three distinct motifs significantly enriched at both time points (Table 5.1). The binding site of zinc finger protein *ZBTB33* (also known as *Kaiso*) was the most significantly enriched motif (M1: E-value $\leq 1.2 \times 10^{-600}$) in *Zfhx3* bound peaks in the SCN, followed by *Thap11* (M2: E-value $\leq 4.4 \times 10^{-217}$) and *CTCF* (M3: E-value $\leq 2.3 \times 10^{-16}$). *ZBTB33* and *Thap11* motifs were only enriched in peaks within promoters as opposed to *CTCF*, which was only enriched in peaks within enhancers. Indeed, previous studies have identified *ZBTB33* to bind highly active promoters (Blattler et al., 2013), while *CTCF* and *Thap11* motifs are commonly found to be significantly enriched within ChIP-seq peaks of various TFs (Worsley Hunt and Wasserman, 2014). However, the *de novo* motif enrichment analysis did not identify the AT rich consensus binding site of *Zfhx3* in the ChIP-seq peaks. The highly enriched *ZBTB33* motif is significantly different to the known *Zfhx3* canonical motifs (Fig. 5.11). This could be attributed to the following scenarios: first, *Zfhx3* primarily functions via indirect binding to the DNA (tethered binding); second, the motifs identified to be enriched (M1, M2 and M3) may be involved in cooperative or competitive binding with *Zfhx3*; or third, the ChIP experiment was unsuccessful to identify *Zfhx3* bound regions (Whitington et al., 2011). Therefore, I further investigated these possibilities and also examined the quality of these motifs.

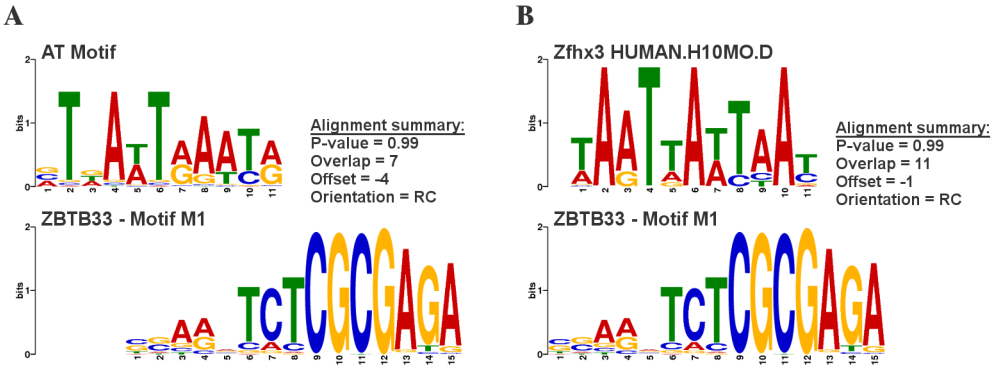


Fig. 5.11 Comparison of *ZBTB33* motif with previously known *Zfhx3* binding motif models. (A) Alignment of *ZBTB33* motif with AT rich *Zfhx3* binding site identified by Parsons et al. (2015). (B) Alignment of *ZBTB33* motif with AT rich *Zfhx3* binding site from HOCOMOCO database. P-value: probability of a same length random motif with equal or better alignment; RC: reverse compliment.

Table 5.1 Motif analysis of *Zfhx3* binding peaks in the SCN.

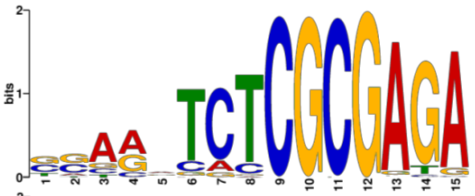
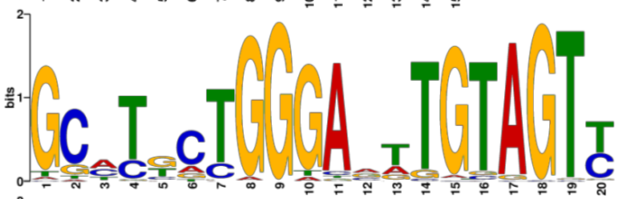
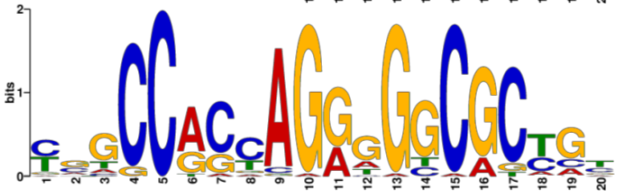
Motif ID	Motif logo	Similar known motif	ZT3			ZT15		
			Top 1k peaks	Promoter: top 1k peaks	Enhancer: top 1k peaks	Top 1k peaks	Promoter: top 1k peaks	Enhancer: top 1k peaks
			E-value (number of peaks)			E-value (number of peaks)		
M1		ZBTB33 (Human)	4.9e-623 (420)	2.3e-600 (409)	Not enriched	1.2e-600 (391)	4.3e-594 (388)	Not enriched
M2		THAP11 (Mouse)	4.1e-241 (334)	3.0e-252 (342)	Not enriched	4.4e-217 (334)	5.3e-259 (345)	Not enriched
M3		CTCF (Mouse)	2.3e-016 (237)	0.032 (233)	2.4e-146 (365)	4.0e-021 (231)	Not enriched	2.4e-123 (333)

Table displays the motif sequences significantly enriched amongst *Zfhx3* binding peaks in the SCN. The motif analysis was performed using three datasets: (1) top 1,000 *Zfhx3* ChIP-seq peaks; (2) top 1,000 promoter associated *Zfhx3* ChIP-seq peaks (< 2 kb upstream and < 200 bp downstream of TSSs); and (3) top 1,000 enhancer associated *Zfhx3* ChIP-seq peaks (> 2 kb upstream and > 200 bp downstream of TSSs). The peaks were analysed using MEME-ChIP (Machanick and Bailey, 2011). The E-value represents the probability of finding the same number of motifs with equal or higher log likelihood ratio in a set of equally sized random sequences.

First, I examined the centrality of the enriched motifs with respect to *Zfhx3* binding peaks. Previous studies have shown that in a successful ChIP-seq experiment, the canonical motif of the ChIP-ed TF is enriched near the centre of the defined TF binding peaks (Bailey and Machanick, 2012). For this purpose, I analysed the distance of the enriched motifs (M1, M2 and M3) with respect to the summit of the peaks using CentriMo (Bailey and Machanick, 2012). In addition to the enriched motifs, the consensus AT motif of *Zfhx3* (from Parsons et al. (2015)) was also analysed. At both time points, the AT motif showed low enrichment and was not centrally enriched (CentriMo corrected $p = 1$) (Fig. 5.12). On the other hand, *ZBTB33* was very centrally enriched with its highest number of motif sites at the centre of the peaks (CentriMo corrected $p < 10^{-104}$). This could indicate two potential scenarios: (1) *Zfhx3* binds to *ZBTB33*, which in turn binds to the DNA (indirect binding); or (2) a non-specific antibody pull down in ChIP-seq. Both possibilities could result in very few peaks with the consensus AT motif. Binding between *Zfhx3* and *ZBTB33* could be a consequence of PPI, however, no evidence of genetic or physical interactions between *Zfhx3* and *ZBTB33* was detected in BioGRID (Stark et al., 2006) and STRING database. On the other hand, *Thap11* and *CTCF* motifs were enriched broadly across the peaks with some enrichment around the peak centres (*Thap11* CentriMo corrected $p = 10^{-11}$; *CTCF* CentriMo corrected $p < 10^{-5}$), which could indicate co-binding activity with *Zfhx3*. However, no evidence of known interactions between *Zfhx3* and *Thap11*, or *Zfhx3* and *CTCF*, was identified in the literature or databases. Overall, these results suggests that either *Zfhx3* binds indirectly to DNA via *ZBTB33*, or the ChIP-seq data also contains regions from potential non-specific interactions.

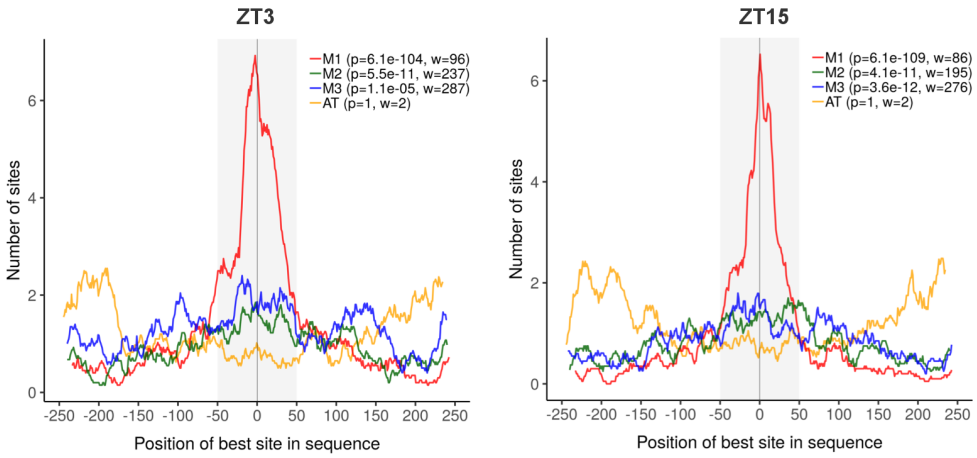


Fig. 5.12 Distribution of motif sites with respect to *Zfhx3* binding peak summits. Distribution plots displaying the number of best scoring motif sites as a function of their relative distance from the summit of the peaks. The density curves were averaged over bins of 20 bp. p: probability that any tested region would be as enriched for best matches to the motif as the reported region; w: width of the most enriched central region; grey shaded area shows ±50 bp region surrounding the peak summit.

Second, I investigated the presence of co-binding between *ZBTB33* (M1), *Thap11* (M2) and *CTCF* (M3) motifs. The motifs occurring within the same ChIP-seq peaks could potentially be indulged in co-binding. To examine this, I compared the binding sites of M1, M2 and M3 in the top 1,000 peaks at ZT3 and ZT15 (Fig. 5.13). Overall, the majority (∼60%) of peaks across both the time points were specific to individual motifs (M1: 28%; M2: 20%; M3:12%), while only ∼18% of the peaks showed co-occurrence between the three motifs (M1+M2+M3: 1.4%; M1+M2: 6.5%; M2+M3: 5.3%; M1+M3: 4.9%) (fraction of peaks were averaged between ZT3 and ZT15). Interestingly, *CTCF* has been previously shown to co-localise and interact with *ZBTB33* (Defossez et al., 2005), which provides evidence for the co-localisation of M1 and M2 motifs. However, no evidence of genetic or physical interactions between *Thap11* and *ZBTB33*, or *Thap11* and *CTCF* were detected in BioGRID and STRING database. Moreover, of the top 1,000 peaks, M1 was detected in only ∼41% of the peaks, while ∼22% of the peaks did not contain any of the top enriched motifs. This shows that the enriched motifs cover only a limited fraction of the top scoring peaks and should be further validated for their prevalence in the remaining peaks.

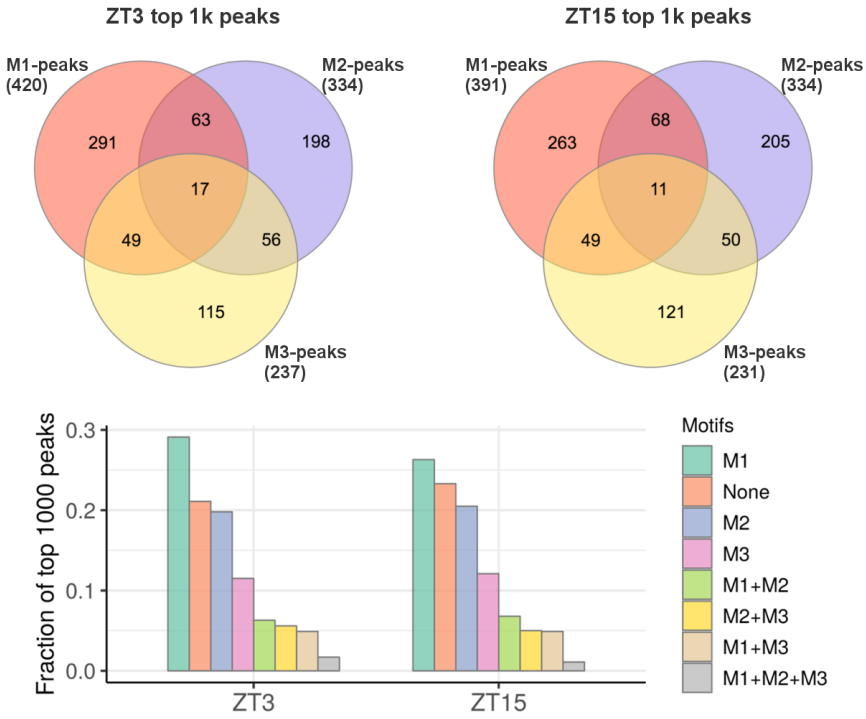


Fig. 5.13 Analysing co-binding between motifs enriched in *Zfhx3* binding peaks. Upper panel: venn-diagrams comparing *Zfhx3* binding peaks associated with different enriched motifs. Lower panel: bar plot displaying the fraction of top 1,000 peaks that are specific and common between the enriched motifs.

Third, I performed a computational validation on the enriched motifs using independent positive control peak sets not used in the motif discovery (training). For each motif, ROC curves were generated which display the motif recognition quality in the positive control sequences compared to random sequences of similar nucleotide composition (see methods 5.3.6). The motifs were validated in three different sets of positive sequences independent from training: (1) top 1,000 peaks not used in training; (2) top 1,000 promoter associated peaks ($\leq$2 kb from a TSS) not used in training; and (3) top 1,000 enhancer associated peaks (> 2 kb from a TSS) not used in training. Since the enriched motifs were present in only a small fraction of the training peaks (as shown earlier), the motifs were also examined in the training sequence set i.e. the top 1,000 peaks which were used for motif discovery.

Overall, all the motifs performed very poorly, even in the training dataset (Fig. 5.14). *Thap11* (M2) was moderately enriched only in the top ~200 peaks of the training dataset (AUC $\leq$ 0.61) and showed poor enrichment in the independent control peaks (AUC $\leq$ 0.56). Likewise, *CTCF* (M3) was only enriched in enhancer control peaks (AUC < 0.66), whereas AT motif did not show enrichment in any of the validation peaks (AUC $\leq$ 0.48). Surprisingly, *ZBTB33* (M1), the most enriched motif identified in the motif discovery analysis, was moderately enriched in the top ~400 training peaks (AUC $\leq$ 0.60), but performed the worst in all the other independent positive control sets (AUC < 0.39). Previous studies have shown *ZBTB33* to bind unmethylated regions associated with highly active promoters (Blattler et al., 2013), which may explain its enrichment only in the top 500 peaks. Indeed, highly active promoters are known to produce false-positives peaks in ChIP-seq experiments (Jain et al., 2015). In order to inspect if *Zfhx3* peaks are associated with highly active promoters, I examined the expression of *Zfhx3* peaks associated genes in the SCN, which revealed that genes associated with *Zfhx3* peaks are likely to be highly expressed compared to genes not associated with the peaks (Fig. 5.15). This suggests that *ZBTB33* motif is likely to be an artefact, perhaps arising from non-specific antibody interactions or poor input control in the ChIP-seq experiment.

Collectively, these results indicate that the enriched motifs (*ZBTB33, Thap11* and *CTCF*) do not follow a common pattern of enrichment throughout the peaks and hence, do not represent a 'genuine' set of binding sites for *Zfhx3*. The motif discovery tool (MEME-ChIP) was sensitive enough to identify motifs enriched in only a fraction of peaks, but these motifs did not validate when tested against independent positive control peaks. Furthermore, the consensus *Zfhx3* AT motif is not enriched in the peaks which could indicate potential non-specific interaction pull-down, or the *Zfhx3* is involved in non-specific binding mediated by a repertoire of different TFs, though it is very hard to distinguish between these scenarios without any additional data.

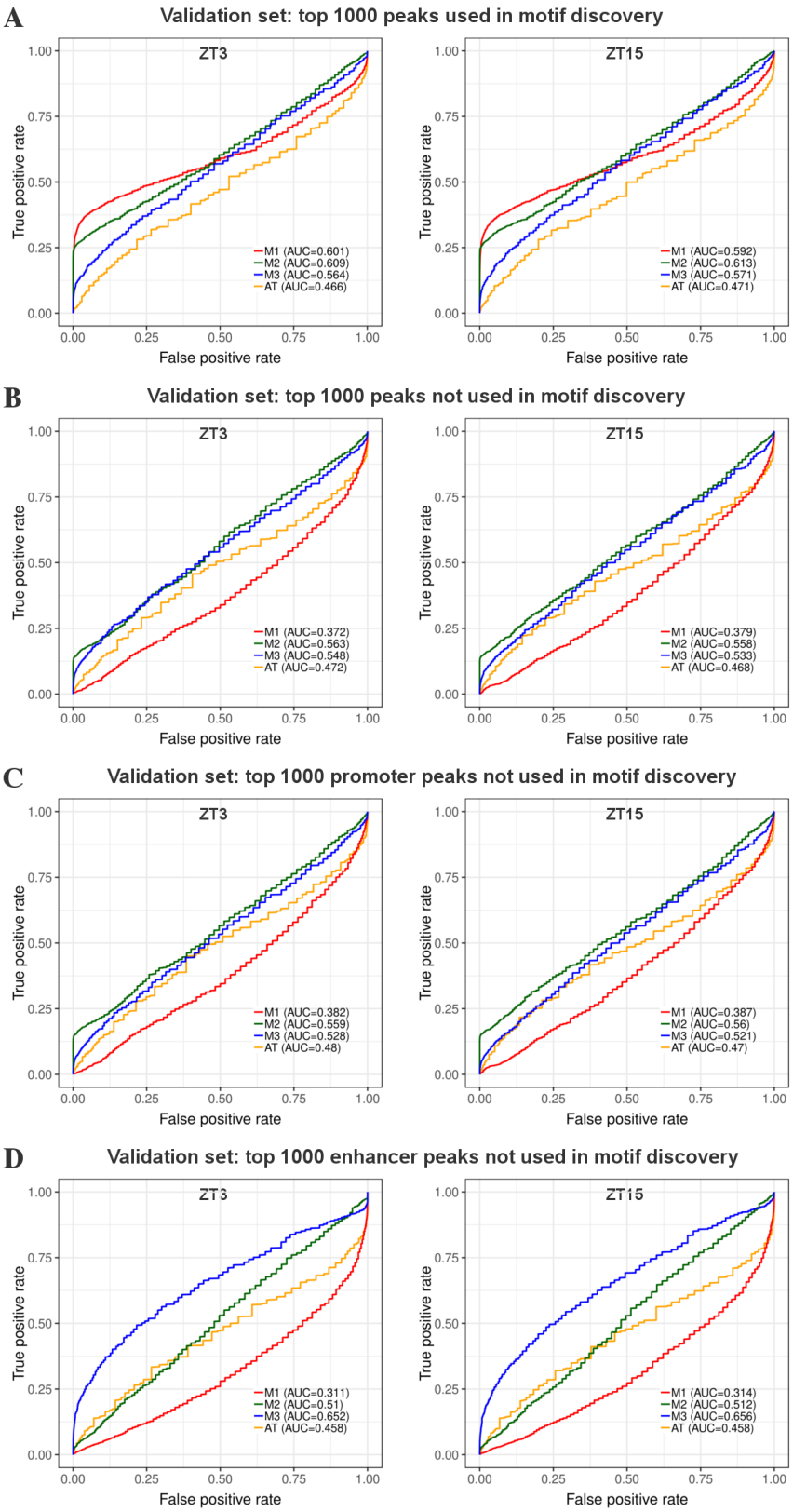


Fig. 5.14 Computational validation assessing the recognition quality of enriched motifs. ROC curves for significantly enriched motifs M1, M2 and M3, along with the AT motif. The curves compare the performance of the motifs in the (A) training dataset and (B-D) different positive control sequences independent of motif discovery. Larger area under the curve represents better motif quality.

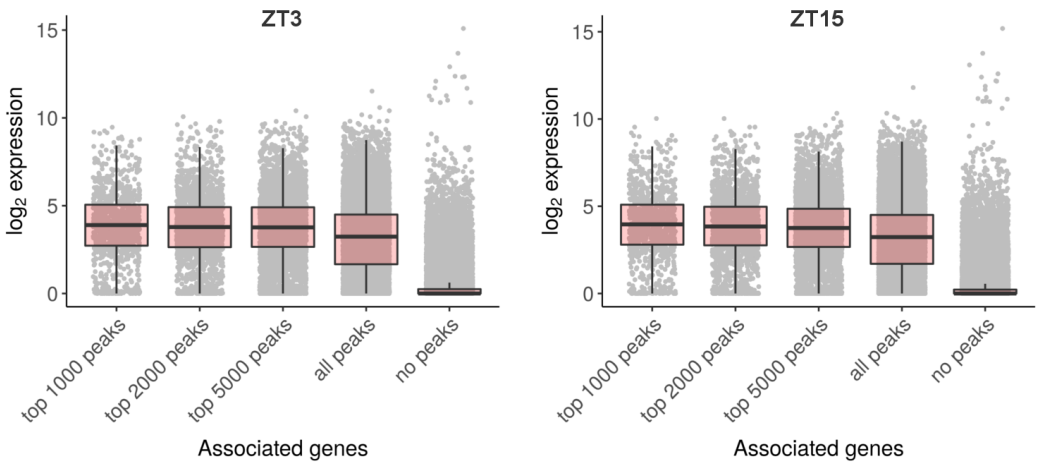


Fig. 5.15 Expression of *Zfhx3* ChIP-seq peaks associated genes in the SCN. Box plots displaying the expression of genes associated with and without *Zfhx3* binding peaks in the SCN. Expression is represented as $\log_2$(RPKM+1).

5.2.6 Comparing *Zfhx3* binding with *Zfhx3*$^{Sci/+}$ transcriptional targets

Finally, I sought to examine the *Zfhx3* ChIP-seq peaks within the differentially expressed genes associated with the *Zfhx3*$^{Sci/+}$ mutation. For this purpose, all the significant ChIP-seq peaks ($q < 0.01$) at ZT3 and ZT15 were associated to potential target genes and combined together into one list, which resulted in 14,222 unique genes. These genes were then compared to the differentially expressed genes associated with the *Zfhx3*$^{Sci/+}$ mutation ($q < 0.05$), which revealed that 72% (121/168) of *Zfhx3*$^{Sci/+}$ transcriptional targets are associated with *Zfhx3* ChIP-seq peaks (Fisher's exact test, $p = 0.012$, OR = 1.54, CI = [1.09, 2.20]) (Fig. 5.16). Of these genes, 54 were identified to be down-regulated and 67 to be up-regulated in the *Zfhx3*$^{Sci/+}$ mice. Furthermore, half of the differentially expressed genes identified in module 1 (11/21) of the *Zfhx3*$^{Sci/+}$ transcriptional network are associated with *Zfhx3* binding peaks (Fisher's exact test, $p = 0.37$, OR = 0.67, CI = [0.25, 1.73]). These genes include neuropeptide *Prok2* and its receptor *Prokr2*, which are involved in circadian intercellular signalling. However, no *Zfhx3* peaks are associated with the neuropeptides *Vip* and *Avp*. Additionally, all the genes defined in module 2 of the *Zfhx3*$^{Sci/+}$ transcriptional network (10/10) are associated with *Zfhx3* binding peaks (Fisher's exact test, $p = 0.016$, OR = Inf, CI = [1.34, Inf]).

Fig. 5.16 Comparison of *Zfhx3* ChIP-seq associated genes with *Zfhx3*$^{Sci/+}$ transcriptional targets. Venn diagram displaying the unique and shared genes associated with *Zfhx3* ChIP-seq peaks and differentially expressed genes in the *Zfhx3*$^{Sci/+}$ mice.

5.3 Methods

5.3.1 Analysis of RNA-seq data

A workflow of the RNA-seq data processing is displayed in Fig. 5.17. The RNA-seq reads were aligned to the mouse genome (mm10) using TopHat (Trapnell et al., 2012). The read counts aligning over each gene were quantified using HTSeq (Anders et al., 2015). The raw read counts in each sample were used to calculate RPKM for each gene. The differentially expressed genes between the $Zfhx3^{+/+}$ and $Zfhx3^{Sci/+}$ mice were identified using three softwares: EdgeR (Robinson et al., 2010), DESeq (Anders and Huber, 2010) and Cufflinks (Trapnell et al., 2012), default parameters were used for all of them. Genes with a q-value (p-value after multiple testing correction) > 0.05 and log2 fold change $< \pm 1$ were filtered out. Genes identified to be differentially expressed by at least two methods were used for populating the heatmap in Fig. 5.3.

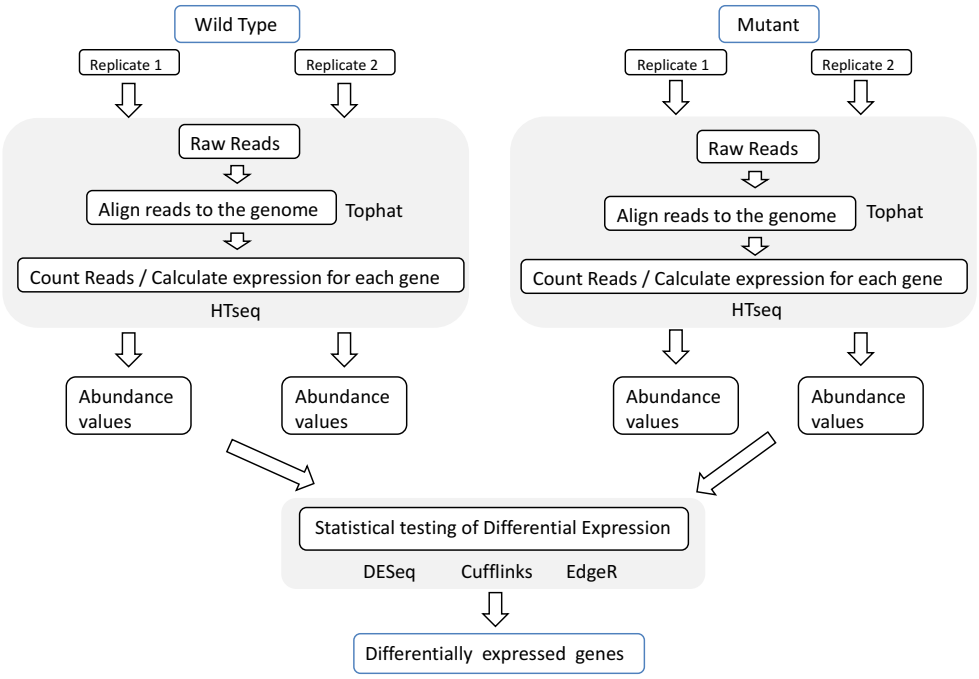


Fig. 5.17 Overview of the RNA-seq pipeline. A schematic representation displaying the processing of RNA-seq data.

5.3.2 PPI and GO enrichment analysis of RNA-seq data

Genes identified to be significantly differentially expressed ($q < 0.05$) between $Zfhx3^{+/+}$ and $Zfhx3^{Sci/+}$ in at least 1 analysis tool were used for network analysis ($n = 168$). PPIs were obtained from the STRING database (Franceschini et al., 2013) and visualised in Cytoscape (Shannon et al., 2003). The MCODE algorithm (Bader and Hogue, 2003)

was used to identify densely connected clusters in the PPI network. The MCODE algorithm is implemented as a Cytoscape plugin. The GO enrichment of genes in each module was performed using g:Profiler (Reimand et al., 2007). The over-represented GO terms were corrected for multiple testing by g:Profiler and terms with q-value < 0.05 were considered significant.

5.3.3 Analysis of AT and other circadian related motifs

Previous studies have identified multiple genes which are directly regulated by *Zfhx3*, such as *Afp, Mrf4, Muc5ac, Pit1* (Berry et al., 2001; Mori et al., 2007; Qi et al., 2008). Parsons et al. (2015) analysed the promoter sequences of these genes to construct a consensus *Zfhx3* binding AT motif using a phylogenetic shadowing based approach. The promoter sequences of these genes were aligned and searched for conserved regions. Using the multiple sequence alignments from UCSC, the AT motif was identified to be conserved across both primates and mammals. The consensus AT motif sequence was built using a mixture model by expectation maximisation (Bailey and Elkan, 1994b). This consensus AT motif was searched in the promoter sequences (450 bp upstream and 50 bp downstream of TSSs) of differentially expressed genes using Pscan (Zambelli et al., 2009) (threshold used: Pscan score > 0.88). Likewise, Pscan was used to identify potential occurrences of the circadian related E-box, D-box and RRE motifs.

5.3.4 Processing of *Zfhx3* ChIP-seq data

The ChIP-seq tags for ZT3 and ZT15, along with that of input control were aligned to the mouse genome version mm10. The ChIP-seq peaks were called using MACS (Model-based Analysis of ChIP-Seq) (Zhang et al., 2008) at a threshold of $q < 0.01$. This identified 27,438 and 27,178 significant peaks at ZT3 and ZT15 respectively. In order to examine what fraction of *Zfhx3* binding in the SCN occurs within active promoters and strong enhancer annotations, the genomic coordinates of these functional segments were compared using BEDTools. The overlap fraction was calculated as: number of base pairs in *Zfhx3* peaks intersecting active promoter or strong enhancer regions, divided by the total genomic coverage of *Zfhx3* peaks in base pairs. The resulting fraction value was then converted into a percentage. To examine if this overlap was significant compared to what one would expect from independent datasets, a permutation test was performed. In each permutation, the genomic coordinates of active promoter/strong enhancer elements (reference data set) were shuffled to random locations ($n = 1{,}000$) preserving the length, chromosome information and the number of elements. The empirical p-value was

calculated as the number of permutations with the overlap fraction greater than the observed overlap fraction, divided by the total number of permutations.

In order to investigate the functional role of *Zfhx3* in the SCN, the top 1,000 ChIP-seq peaks at ZT3 and ZT15 were associated to potential target genes using GREAT. In cases where GREAT predicted multiple target genes for a particular *Zfhx3* binding peak, the gene nearer to the peak was selected as the primary predicted target. GREAT was run using default parameters on mm10 assembly and the whole genome was selected for control background regions. This approach identified 977 and 986 potential *Zfhx3* gene targets at ZT3 and ZT15 respectively. These gene lists were combined together resulting in 1,216 unique genes. To examine the functional associations of these genes, GO enrichment analysis was performed using g:Profiler. Moderate filtering within g:Profiler was applied on significant GO terms (corrected p-value < 0.05) to obtain the best GO term per parent.

In order to detect changes in the binding affinity of *Zfhx3* binding between ZT3 and ZT15, differential binding analysis was performed using DiffBind (Ross-Innes et al., 2012). A threshold of $p < 0.05$ and fold change $> \pm 2$ was applied to the results, which detected no significant differentially bound peaks between ZT3 and ZT15.

5.3.5 Motif analysis of *Zfhx3* ChIP-seq peaks

To identify the *Zfhx3* binding motif and other enriched motifs within the ChIP-seq peaks, MEME-ChIP (Machanick and Bailey, 2011) was used to conduct a *de novo* motif analysis. Only the top 1,000 peaks (based on q-value) were used for this analysis, as the top scoring peaks tend to have the highest binding affinity and the least background noise. For each peak, genomic sequence within ± 250 bp from the summit of the peak was extracted. MEME-ChIP first detects the enriched motif sequences amongst the set of ChIP-seq peaks and then matches them to previously known motif models. Additionally, I examined the pattern of enriched motifs in peaks within promoter regions and enhancer (distal) regions. For this purpose, the *Zfhx3* peaks were classified into either promoter or enhancer associated on the basis of the following rule: peaks binding within 2 kb upstream and 200 bp downstream of a known TSS were considered to be promoter associated, while the remaining peaks (> 2 kb upstream, > 200 bp downstream) were considered to be enhancer associated. This resulted in 17,673 and 17,505 promoter associated peaks at ZT3 and ZT15 respectively ($\sim$64% at both time points). The top 1,000 peaks (based on q-value) within promoter and enhancer regions were then analysed for motif enrichment.

Next, the PWM models of the top enriched motifs were used as input for FIMO (Grant et al., 2011) to identify the potential motif sites in the top 1,000 peaks at ZT3

and ZT15. Only the motif matches attaining a FIMO p-value $< 10^{-4}$ were considered, where the p-value represents the probability of a random sequence of the same length as the motif matching that position of the sequence with as good as or a better score. To examine the co-binding between the enriched motifs, the genomic locations of each motif derived from FIMO were mapped to the peaks. The peak IDs were then compared between the enriched motifs to identify potential motif sites occurring within the same peaks. The location of the motifs were further analysed with respect to the summit of the peaks using CentriMo (Bailey and Machanick, 2012). The consensus AT motif identified by Parsons et al. (2015) was also analysed in a similar way.

5.3.6 Assessing the recognition quality of the enriched motifs

To examine the authenticity of the enriched motifs, a computational validation of the motif models (PWMs) was performed using the strategy described in (Kulakovskiy et al., 2013). In addition to the enriched motifs, the AT motif was also analysed to examine whether it is enriched in the validation sequence set. Each PWM was first validated on the training set (sequences/peaks used for motif discovery) itself, and then using independent positive control sequences not used in the motif discovery process. For true-positive cases, peaks from the positive set with at least one PWM match scoring equal or better than the threshold were used.

For each PWM, the positive control sequences were sorted based on their decreasing PWM match scores. This decreasing set of PWM scores were considered as PWM threshold values, where each threshold corresponded to a true-positive rate value. Next, for each PWM threshold, the likelihood P_s of identifying at least one PWM match with an equal or better score than threshold, in a random double-stranded DNA segment of length L. The length of the sequences in the positive control set was used as L (i.e. 500). For a motif of length l, P_s was calculated as:

$$P_s = 1 - (1 - P)^{2(L-l+1)} \tag{5.1}$$

where P is the probability of getting a given score for a random word at the particular position of a random double-strand DNA sequence (as computed in Touzet and Varré (2007)), expecting the matches (including overlapping matches) are independent and their number abide by compound poisson distribution.

To evaluate and visualise the recognition quality of the PWMs, ROC curves were plotted displaying the true-positive rate as a function of P_s (which is considered as an estimation of the false-positive rate) for a set of PWM thresholds based on the positive

control sequences. To quantify and compare the PWM quality, AUC was computed. A higher AUC value corresponds to better motif recognition quality.

5.4 Discussion

The core circadian TTFL circuitry in mammals comprises of a network of TFs, regulatory co-factors and genes that regulate the intrinsic molecular clocks. During the last decade, studies have been mostly focused on the characterisation of core circadian genes which have uncovered their significance in maintaining the 24 hour cycles of internal molecular clocks. Here, we provide evidence of *Zfhx3* as a novel gene beyond the core TTFL network that operates a clock-regulated transcriptional axis, possibly by controlling a complex of neuropeptides responsible for molecular clock synchrony in individual SCN cells. This study extends the catalogue of regulatory proteins engaged in sustaining stable circadian rhythms in mammals.

With the increasing popularity of RNA-seq, multiple computational approaches and software packages have been developed to analyse and interpret RNA-seq data. However, none of the methods are optimal for all types of experimental data and can produce results with large variation (Seyednasrollah et al., 2015). In order to overcome such problems, I developed an analysis pipeline integrating multiple approaches to identify differential expression, which could reduce false positives and capture a wide range of genes potentially contributing to the phenotype. This approach detected that the majority of differentially expressed genes in *Zfhx3*$^{Sci/+}$ mice were down-regulated, including expression of circadian related neuropeptides and receptors. Moreover, scanning the promoters of differentially expressed genes predicted 39% of them to contain the consensus AT motif. These genes may be directly regulated by *Zfhx3*, while the remaining differentially expressed genes may have altered expression possibly as a result of indirect regulation by *Zfhx3* in the SCN. This indicates that other unknown factors are also involved in the maintenance of stable circadian oscillations in SCN. Although not explored yet, identification of these *Zfhx3* binding partners would further aid in understanding its function in the SCN. Another aspect which requires further investigation is the complex of ribosomal proteins (module 2) up-regulated in *Zfhx3*$^{Sci/+}$ mice. This observation indicates that *Zfhx3* may indirectly contribute to the structural integrity of ribosomal subunits via regulating ribosomal proteins, a *Zfhx3* function not known yet.

Previous studies have shown that *Zfhx3* has the ability to both activate (Qi et al., 2008) and suppress (Mori et al., 2007; Yasuda et al., 1994) the expression of its target genes via the AT motif. Our data shows that *Zfhx3* can activate the expression of a

network of neuropeptides via AT motif binding, and that this function is distorted in *Zfhx3*$^{Sci/+}$ mice. Some *Zfhx3*$^{Sci/+}$ phenotypes are common with features observed in previous studies involving mouse models associated with neuropeptides. For instance, mice not producing the *Vip* protein show arrhythmic or reduced circadian period relative to wild type mice (Aton et al., 2005; Brown et al., 2007). Likewise, mouse models lacking *Prok2* and *Prokr2* display a reduction in the amplitude of locomotor activity (Li et al., 2006; Prosser et al., 2007). Overall, these observations suggest that the reduced ability of *Zfhx3* to transcriptionally activate the neuropeptide hub in *Zfhx3*$^{Sci/+}$ mice contributes to the observed circadian phenotype. The findings from this study further propose the involvement of AT motif in synchronous cellular rhythms which has not been previously found. This AT motif recognised and bound by the TF *Zfhx3* now expands the group of DNA motifs involved in the regulation of circadian function.

ChIP-seq has been widely used to identify the *in vivo* binding sites of TFs, which can further help to investigate the regulatory network of the TF of interest. Using ChIP-seq, the *Zfhx3* binding peaks were mapped in the SCN at ZT3 and ZT15, revealing *Zfhx3* preferentially binds to promoter regions. The ChIP-seq analysis identified no significant difference in binding affinity of *Zfhx3* between ZT3 and ZT15. The *Zfhx3* binding peaks were detected to be enriched in *ZBTB33, Thap11* and *CTCF* motifs, however, these motifs were present in only a small fraction of the peaks and showed very little enrichment in positive control peaks independent of motif discovery. This suggests that the enriched *ZBTB33, Thap11* and *CTCF* motifs may not represent the 'real' motif sequence bound by *Zfhx3*. Prior research has also reported *CTCF* and *Thap11* motifs to be frequently over-represented within ChIP-seq peaks of various TFs (Worsley Hunt and Wasserman, 2014). Surprisingly, the *Zfhx3* bound peaks showed no evidence for the enrichment of the consensus AT motif. This could suggest that *Zfhx3* does not directly bind to the DNA and instead rides on other co-regulators, however, almost all the inspected peaks were completely devoid of AT motif which is doubtful. It remains unclear whether these ChIP-seq regions are associated to immunoprecipitated *Zfhx3*, or if the majority of these regions are a result of some other cause such as non-specific interactions or poor quality of the standard input control, and hence should be considered as false-positives until further validated.

Further inspection of the ChIP-seq peaks showed the possibility of a common false positive signal in this data, which occurs in many ChIP-seq profiles. Previous studies have shown that despite using an antibody validated for its specificity, highly open chromatin regions with high TF and co-factor occupancy can provide interaction-prone surfaces which could result in non-specific interactions of the antibody (Jain et al., 2015). These false-positive ChIP-seq regions, referred to as 'Phantom Peaks', are significantly associated with highly active promoters and regions bound by large number of TFs.

In the case of *Zfhx3* ChIP-seq data, the majority (∼70%) of binding is identified in promoter regions (within 3 kb of TSSs). Indeed, ∼59% of the genomic regions covered by *Zfhx3* peaks overlap with active promoters in the cerebellum, cortex or whole brain. Furthermore, previous research has shown that *ZBTB33* binds to promoters of highly expressed genes which correlates with the high expression levels of genes associated with the top 1,000 *Zfhx3* ChIP-seq peaks in the SCN (Fig. 5.15). These observations indicate the possible presence of 'Phantom Peaks' in this dataset. However, it is difficult to confirm this without any additional data. Ideally, the ChIP-seq peaks in this data could be validated by performing another ChIP-seq experiment in the SCN cells lacking *Zfhx3* protein. If the majority of the binding peaks are retained, that would prove the observed peaks are false-positives or vice versa.

Chapter 6

Summary and future directions

The data presented in this thesis has helped us to understand the relationship between enhancers and gene function in the mouse genome. However, there are several questions which remain unanswered. A significant portion of the thesis focussed on identifying potential enhancers in the mouse genome and investigating their functional properties. Whereas, the second and the last chapter of the thesis focussed on MRCHI specific mouse models where the aim was to decipher the regulatory networks associated with the altered TFs in the mutant mouse. The overall purpose of this thesis was to gain more understanding about the enhancer regulatory networks in the mouse genome, and investigate how their presence may affect gene expression and phenotypes.

Initially, I analysed the *Klf14* associated transcriptional targets in a *Klf14* knockout mouse model, with the aim to compare them with the *KLF14 trans*-network in human. The results show that despite a conserved *KLF14* regulatory motif between the human and mouse genomes, *KLF14* transcriptional targets are mostly species specific. However, there are significant protein-protein interactions between the transcriptional targets in the two species, which show their potential involvement in the same functional pathway. A further ChIP-seq experiment in the mouse to identify *Klf14* binding would be useful to accurately reveal its regulatory targets across the mouse genome and its associated regulatory pathways. The *KLF14* associated adipose-specific enhancer in humans which harbours binding sites for adipogenesis TFs and T2D risk variants, is likely to have no regulatory activity in the mouse. I suggest that the loss of this enhancer in the mouse genome may have shifted the function of *Klf14* largely towards other areas such as cholesterol metabolism, and therefore we observe no T2D associated phenotypes in the *Klf14* knockout mouse model. This study is a good example depicting the usefulness of such enhancer annotations for comparing regulatory mechanisms across species and also in translating biological knowledge between mouse models and humans.

What T2D associated SNPs in the *KLF14* locus are *cis*-regulatory variants and what TFBS they disrupt are interesting but unanswered questions. Genetic variants associated with diseases are being identified at a rapid pace through GWASs. The variants occurring in enhancers, especially within the TFBSs, may disrupt the regulatory networks of their associated TFs, leading to the loss of a normal healthy state. Based on the density of conserved TFBSs near the T2D associated SNPs in the *KLF14* locus, I identified 16 SNPs which could be potential *cis*-regulatory variants, but could not detect any common TFBS they disrupt. Future work should focus on developing improved computational methods like PMCA, to more accurately detect the TF binding network being altered by the *cis*-regulatory variants. Integrating methods like PMCA with ChIP-seq profiles, DHSs and histone modification data would help in detecting functionally active TFBSs with greater accuracy. Such methods would be immensely useful to functionally characterise these genetic variants occurring within the regulatory regions.

The advancement in sequencing technologies has made it possible to capture a detailed snapshot of enhancer profiles at a genome-wide scale. In this thesis, I have produced a catalogue of multiple enhancer types in a diverse range of mouse tissues and cell-types. This catalogue includes well defined tissue-specific enhancers, super-enhancers (SEs), typical-enhancers (TEs) and weak-enhancers in previously unexplored tissues. To produce these enhancer annotations, I used ChIP-seq data from only three histone marks, primarily because they were the most investigated histone marks at the time of producing this dataset, and hence, their ChIP-seq data was available for all the 22 mouse tissues analysed here. Future work would involve improving these annotations by using more histone marks and additional datasets such as DHSs. The tissue-specific enhancers in this catalogue were identified using the Tau metric which performed better than the previously applied clustering methods. However, I used previously defined thresholds for the Tau score. Future studies should optimise the thresholds for the Tau score in such a way that it enables us to capture the maximum number of tissue-specific elements with minimum amount of noise. Furthermore, adding more tissues and cell-types to this catalogue in the future would further refine the quantification of tissue-specific enhancers.

Which regions in the mouse genome should be targeted to functionally characterise disease-associated SNPs (DA-SNPs) from GWASs is an important question. In this thesis, I have shown that non-coding DA-SNPs from some GWAS traits were enriched in mouse enhancers of disease-relevant tissues. However, this analysis was performed using a small curated set of non-coding DA-SNPs from only 26 GWAS traits. A similar, but more comprehensive study incorporating all the DA-SNPs and disease traits in GWAS catalogue (Welter et al., 2014) should be conducted as this may help in predicting the disease traits that can be better replicated in mouse models.

I explored how enhancers influence the expression and phenotypes of their target genes in order to understand gene regulation in cells. By comparing SEs and TEs, I investigated their distinctive roles in gene function, as many researchers remain dubious about such sub-categorisation of strong enhancers. SEs were identified to drive high total-expression and tissue-specific expression of their associated genes compared to TEs. However, a major finding in this thesis was that SE and TE associated genes share common phenotypic outcomes even though their expression profiles and overall numbers in the genome differ. The evidence described in this thesis show that there is no significant difference in severity and breadth of phenotypes produced from the knockouts of SE and TE associated genes. Following on from this, another interesting finding was the high number of genes identified to be associated with TEs, which highlights their importance in the genome - an observation consistent with a recent report (Hamdan and Johnsen, 2018). While the majority of the previous studies have focussed on SEs mainly because of their association with key cell identity genes, other strong enhancers in the genome also appear to have a notable influence on the gene function.

Irrespective of enhancer classes, the results described in this thesis increases our understanding about the correlation between enhancers and gene function, which could help in predicting the effect of enhancer disruption. However, it is important to note that the enhancers were indirectly associated to phenotypes via the phenotype outcomes of their potential target genes. An assessment of direct association between enhancers and phenotypes would be difficult to perform *in silico*, and would require *in vivo* enhancer-loss of function studies to accurately identify its phenotypic associations. Similar studies involving deletion of enhancers using CRISPR-Cas9 have been performed in the past (Groschel et al., 2014; Hay et al., 2016; Li et al., 2014; Moorthy et al., 2017; Shin et al., 2016). However, such *in vivo* functional characterisation of enhancers rely on comprehensive and well defined enhancer catalogues, such as the one produced in this thesis. The wide range of tissues and cell-types analysed in this thesis provides a platform to further characterise enhancer regulatory mechanisms in the mouse genome. These studies would be helpful in developing novel approaches to manipulate gene expression in research experiments and in drug development, for example BET inhibitors are being used to disrupt SE activity which causes a reduction in the expression of oncogenes (Drier et al., 2016; Loven et al., 2013; Peeters et al., 2015; Pelish et al., 2015; Wyce et al., 2013).

Another important question which this thesis addressed is whether the constituent enhancers within SEs or TEs exert an additive or a more complex cooperative effect on target gene expression. The analysis described in this thesis predicts that total- and tissue-specific expression levels are weakly correlated with the number of constituent

enhancers at a genome-wide scale. Therefore, all constituent enhancers do not appear to contribute to the transcriptional output with the same strength suggesting a non-additive relationship between them. However, there is a possibility that some constituent enhancers are redundant and do not contribute to the transcriptional output in normal conditions (Cannavo et al., 2016; Frankel et al., 2010; Hong et al., 2008; Moorthy et al., 2017). It is a difficult task to distinguish redundant enhancers from active enhancers *in silico*, especially with the dataset collected and the computational nature of analysis conducted in this thesis. Further *in vivo* experiments involving single and combinatorial deletions of such constituent enhancers would be required to accurately detect their function in gene expression. Though similar studies have been performed to decipher the function of individual SE constituents towards expression of a single or a group of genes (Hay et al., 2016; Hnisz et al., 2015; Moorthy et al., 2017; Shin et al., 2016; Suzuki et al., 2017), their findings are highly variable. Furthermore, only a few genomic loci have been investigated by such experiments and we cannot extrapolate the findings from a few loci to the whole genome. Clearly, more studies are required as the question still remains unanswered about the relationship between individual enhancer elements at the genome-wide scale. Do these enhancers function independently? Do they interact with each other and influence each other's activity? The answers to these questions will be important to understand whether SEs represent a new class of regulatory elements and function as a single unit, or they are simply a cluster of traditional enhancers.

Since I used histone modification ChIP-seq data to predict genome-wide enhancers, I could only predict a coarse level of enhancer annotations due to the limitations associated with it. Histone modification data is good for predicting enhancers because it allows us to effectively distinguish enhancers from other regulatory activity in the genome. However, histone modifications occur at the regions flanking regulatory elements, hence producing a broad ChIP-seq signal around regulatory regions like enhancers. These extended signals can often introduce noise in the process of chromatin state segmentation resulting in: (1) false enhancer regions; or (2) enhancer regions with inaccurate boundaries and overestimated length. Therefore, future studies should focus on using open chromatin signals captured by technologies such as DNase-seq or ATAC-seq, which have been observed to predict enhancer regions at a finer resolution. These datasets if applied here would generate enhancer annotations with increased genomic location accuracy and fewer false-positives, which in turn would refine the classes of enhancer associated genes and produce more significant enrichment results. As most of the downstream analysis conducted in this thesis to investigate enhancer associated functions were based on enrichment whilst seeking common patterns, implementing DNase-seq/ATAC-seq data would produce similar enrichment outcomes and hence not affect the conclusions made in this thesis. However, as a result of less noise and fewer false-positives in the data, analysis which was not based on enrichment could be

benefited, in particular: (1) the correlation between the number of constituent enhancers and target gene expression may improve; (2) the machine learning features associated with tissue-specific enhancers/promoters in the random forest classifier may have a cleaner signal, which in turn may improve the predictive capacity of tissue-specific regulatory elements to infer gene-phenotype associations.

Tremendous progress has been made in the last five years in the field of enhancer discovery. As studies similar to this thesis produce more characterised enhancer annotations, the next step in the future is to functionally characterise them, either through massively parallel reporter assays (Kwasnieski et al., 2012; Melnikov et al., 2012), or through high-throughput CRISPR-Cas9 based functional screening (Canver et al., 2015; Diao et al., 2016; Korkmaz et al., 2016; Peeters et al., 2015). It would be essential to test these elements in their natural chromatin environment, however, it would be challenging as the enhancers detected to have no activity in their natural chromatin context may have a subtle effect, which may be difficult to capture using the current phenotypic tests, or enhancers may have a critical role at a different stage of development. These functional studies could play a key role in testing enhancer malfunction in diseases by generating custom alleles and evaluating their chromatin interactions. It would also be important to understand how the presence or absence of genetic variants within enhancer regions affect drug responses. However, several questions remain outstanding: which elements are critical to maintain the function of enhancers? Whether the distance between the individual constituent enhancers have any effect on their function? How is the regulatory information encoded in the DNA sequence? What controls the mechanisms that pair enhancer with their target genes?

Lastly, I investigated the regulatory network altered by the *Zfhx3*$^{Sci/+}$ mutation in a MRCHI circadian mouse model. The evidence provided in this thesis showed that the *Zfhx3*$^{Sci/+}$ mutation disrupts the ability of *Zfhx3* to activate transcription of a circadian related neuropeptide network which contributed to the *Sci* phenotype. This study identified *Zfhx3* as a novel gene beyond the TTFL network to be involved in controlling the circadian oscillations. However, not all the genes affected by the *Zfhx3*$^{Sci/+}$ mutation were identified to be directly regulated by *Zfhx3* via an AT motif, which suggests the involvement of other unknown factors or motifs in the functional pathway contributing to the circadian phenotype. Future analysis would involve the identification of potential co-factors or *Zfhx3* binding partners in this network. This could be primarily achieved by performing a *de novo* motif analysis amongst the *Zfhx3*$^{Sci/+}$ transcriptional targets to identify other TF binding motifs enriched within these genes, and inspecting if the location of these motifs is near the AT motif. I characterised the genes altered by the *Zfhx3*$^{Sci/+}$ mutation into different functional modules. The first module consisted of a network of neuropeptides known to be involved in circadian rhythms, while another

module consisted of ribosomal proteins (module 2). The second module was not explored, therefore, future analysis should also focus on investigating the function of *Zfhx3* in regulating ribosomal proteins.

Overall, in this thesis I have shown how the presence of enhancers affect the gene function in the mouse genome. All this evidence goes to prove that they are important contributors in disease causation, and understanding the mechanisms behind their function could lead to clues on disease pathogenesis and possible therapies.

List of publications

Parsons, M. J., M. Brancaccio, **S. Sethi**, E. S. Maywood, et al. (2015). "The Regulatory Factor ZFHX3 Modifies Circadian Function in SCN via an AT Motif-Driven Axis". In: *Cell* 162.3, pp. 607-621. ISSN: 0092-8674. DOI: 10.1016/j.cell.2015.06.060.

Balzani, E., G. Lassi, S. Maggi, **S. Sethi**, M. J. Parsons, M. M. Simon, P. M. Nolan, V. Tucci (2016). "The Zfhx3-Mediated Axis Regulates Sleep and Interval Timing in Mice". In: *Cell Rep* 16.3, pp. 615-21. DOI: 10.1016/j.celrep.2016.06.017.

Potter, P. K., M. R. Bowl, P. Jeyarajan, L. Wisby, [and 49 others including **S. Sethi**] (2016). "Novel gene function revealed by mouse mutagenesis screens for models of age-related disease". In: *Nat Commun* 7, p. 12444. ISSN: 2041-1723. DOI: 10.1038/ncomms12444.

Small, K. S., M. Todorčević, M. Civelek, J. S. El-Sayed Moustafa, X. Wang, M. S. Simon, J. Fernandez-Tajes, A. Mahajan, M. Horikoshi, A. Hugill, C.A. Glastonbury, L. Quaye, M. J. Neville, **S. Sethi**, et al. (2018). "Regulatory variants at KLF14 influence type 2 diabetes risk via a female-specific effect on adipocyte size and body composition". In: *Nature Genetics* 50.4, pp. 572-580. ISSN: 1546-1718. DOI: 10.1038/s41588-018-0088-x.

Sethi, S., I. E. Vorontsov, I. V. Kulakovskiy, S. Greenaway, J. Williams, V. J. Makeev, S. D. M. Brown, M. M. Simon, A.-M. Mallon (2019)."Deciphering the impact of enhancer architecture on gene function and mouse phenotypes". Under review in *Cell Reports*.

References

Achour, M., S. Le Gras, C. Keime, F. Parmentier, et al. (2015). "Neuronal identity genes regulated by super-enhancers are preferentially down-regulated in the striatum of Huntington's disease mice". In: *Hum Mol Genet* 24.12, pp. 3481–96. ISSN: 0964-6906. DOI: 10.1093/hmg/ddv099.

Adam, R. C., H. Yang, S. Rockowitz, S. B. Larsen, et al. (2015). "Pioneer factors govern super-enhancer dynamics in stem cell plasticity and lineage choice". In: *Nature* 521.7552, pp. 366–70. ISSN: 0028-0836. DOI: 10.1038/nature14289.

Adams, M. D., S. E. Celniker, R. A. Holt, C. A. Evans, et al. (2000). "The genome sequence of Drosophila melanogaster". In: *Science* 287.5461, pp. 2185–95. ISSN: 0036-8075 (Print) 0036-8075.

Ahituv, N., Y. Zhu, A. Visel, A. Holt, et al. (2007). "Deletion of ultraconserved elements yields viable mice". In: *PLoS Biol* 5.9, e234. ISSN: 1544-9173. DOI: 10.1371/journal.pbio.0050234.

Ahmadiyeh, N., M. M. Pomerantz, C. Grisanzio, P. Herman, et al. (2010). "8q24 prostate, breast, and colon cancer risk loci show tissue-specific long-range interaction with MYC". In: *Proc Natl Acad Sci U S A* 107.21, pp. 9742–6. ISSN: 0027-8424. DOI: 10.1073/pnas.0910668107.

Akazawa, H. and I. Komuro (2003). "Roles of Cardiac Transcription Factors in Cardiac Hypertrophy". In: *Circulation Research* 92.10, p. 1079.

Almer, A., H. Rudolph, A. Hinnen, and W. Horz (1986). "Removal of positioned nucleosomes from the yeast PHO5 promoter upon PHO5 induction releases additional upstream activating DNA elements". In: *Embo j* 5.10, pp. 2689–96. ISSN: 0261-4189 (Print) 0261-4189.

Almgren, P., M. Lehtovirta, B. Isomaa, L. Sarelin, et al. (2011). "Heritability and familiality of type 2 diabetes and related quantitative traits in the Botnia Study". In: *Diabetologia* 54.11, pp. 2811–9. ISSN: 0012-186x. DOI: 10.1007/s00125-011-2267-5.

Anders, S. and W. Huber (2010). "Differential expression analysis for sequence count data". In: *Genome Biol* 11.10, R106. ISSN: 1474-7596. DOI: 10.1186/gb-2010-11-10-r106.

Anders, S., P. T. Pyl, and W. Huber (2015). "HTSeq–a Python framework to work with high-throughput sequencing data". In: *Bioinformatics* 31.2, pp. 166–9. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btu638.

Angelis, M. H. de, G. Nicholson, M. Selloum, J. White, et al. (2015). "Analysis of mammalian gene function through broad-based phenotypic screens across a consortium of mouse clinics". In: *Nat Genet* 47.9, pp. 969–978. ISSN: 1061-4036. DOI: 10.1038/ng.3360.

Antoniou, M., E. deBoer, G. Habets, and F. Grosveld (1988). "The human beta-globin gene contains multiple regulatory regions: identification of one promoter and two downstream enhancers". In: *Embo j* 7.2, pp. 377–84. ISSN: 0261-4189 (Print) 0261-4189.

Arnold, C. D., D. Gerlach, C. Stelzer, Ł. M. Boryń, et al. (2013). "Genome-Wide Quantitative Enhancer Activity Maps Identified by STARR-seq". In: *Science* 339.6123, pp. 1074–1077.

Arnone, M. I. and E. H. Davidson (1997). "The hardwiring of development: organization and function of genomic regulatory systems". In: *Development* 124.10, pp. 1851–64. ISSN: 0950-1991 (Print) 0950-1991.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, et al. (2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium". In: *Nat Genet* 25.1, pp. 25–9. ISSN: 1061-4036 (Print) 1061-4036. DOI: 10.1038/75556.

Aton, S. J., C. S. Colwell, A. J. Harmar, J. Waschek, et al. (2005). "Vasoactive intestinal polypeptide mediates circadian rhythmicity and synchrony in mammalian clock neurons". In: *Nat Neurosci* 8.4, pp. 476–83. ISSN: 1097-6256 (Print) 1097-6256. DOI: 10.1038/nn1419.

Axel, R., H. Cedar, and G. Felsenfeld (1973). "Synthesis of globin ribonucleic acid from duck-reticulocyte chromatin in vitro". In: *Proc Natl Acad Sci U S A* 70.7, pp. 2029–32. ISSN: 0027-8424 (Print) 0027-8424.

Ayadi, A., M. C. Birling, J. Bottomley, J. Bussell, et al. (2012). "Mouse large-scale phenotyping initiatives: overview of the European Mouse Disease Clinic (EUMODIC) and of the Wellcome Trust Sanger Institute Mouse Genetics Project". In: *Mamm Genome* 23.9-10, pp. 600–10. ISSN: 0938-8990. DOI: 10.1007/s00335-012-9418-y.

Babu, M. M., N. M. Luscombe, L. Aravind, M. Gerstein, et al. (2004). "Structure and evolution of transcriptional regulatory networks". In: *Curr Opin Struct Biol* 14.3, pp. 283–91. ISSN: 0959-440X (Print) 0959-440x. DOI: 10.1016/j.sbi.2004.05.004.

Bader, G. D. and C. W. Hogue (2003). "An automated method for finding molecular complexes in large protein interaction networks". In: *BMC Bioinformatics* 4, p. 2. ISSN: 1471-2105.

Bailey, T. L. and C. Elkan (1994a). "Fitting a mixture model by expectation maximization to discover motifs in biopolymers". In: *Proc Int Conf Intell Syst Mol Biol* 2, pp. 28–36. ISSN: 1553-0833 (Print) 1553-0833.

— (1994b). "Fitting a mixture model by expectation maximization to discover motifs in biopolymers". In: *Proc Int Conf Intell Syst Mol Biol* 2, pp. 28–36. ISSN: 1553-0833 (Print) 1553-0833.

Bailey, T. L. and P. Machanick (2012). "Inferring direct DNA binding from ChIP-seq". In: *Nucleic Acids Res* 40.17, e128. ISSN: 0305-1048 (Print). DOI: 10.1093/nar/gks433.

Banerji, J., L. Olson, and W. Schaffner (1983). "A lymphocyte-specific cellular enhancer is located downstream of the joining region in immunoglobulin heavy chain genes". In: *Cell* 33.3, pp. 729–40. ISSN: 0092-8674 (Print) 0092-8674.

Banerji, J., S. Rusconi, and W. Schaffner (1981). "Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences". In: *Cell* 27.2 Pt 1, pp. 299–308. ISSN: 0092-8674 (Print) 0092-8674.

Barbieri, C. E., S. C. Baca, M. S. Lawrence, F. Demichelis, et al. (2012). "Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer". In: *Nat Genet* 44.6, pp. 685–9. ISSN: 1061-4036. DOI: 10.1038/ng.2279.

Bass, J. and M. A. Lazar (2016). "Circadian time signatures of fitness and disease". In: *Science* 354.6315, pp. 994–999. ISSN: 0036-8075. DOI: 10.1126/science.aah4965.

Bassuny, W. M., K. Ihara, Y. Sasaki, R. Kuromaru, et al. (2003). "A functional polymorphism in the promoter/enhancer region of the FOXP3/Scurfin gene associated with type 1 diabetes". In: *Immunogenetics* 55.3, pp. 149–156. ISSN: 0093-7711 (Print) 0093-7711. DOI: 10.1007/s00251-003-0559-8.

Bauer, D. E., S. C. Kamran, S. Lessard, J. Xu, et al. (2013). "An erythroid enhancer of BCL11A subject to genetic variation determines fetal hemoglobin level". In: *Science* 342.6155, pp. 253–7. ISSN: 0036-8075. DOI: 10.1126/science.1242088.

Bejerano, G., M. Pheasant, I. Makunin, S. Stephen, et al. (2004). "Ultraconserved Elements in the Human Genome". In: *Science* 304.5675, p. 1321.

Belton, J. M., R. P. McCord, J. H. Gibcus, N. Naumova, et al. (2012). "Hi-C: a comprehensive technique to capture the conformation of genomes". In: *Methods* 58.3, pp. 268–76. ISSN: 1046-2023. DOI: 10.1016/j.ymeth.2012.05.001.

Benton, M. L., S. C. Talipineni, D. Kostka, and J. A. Capra (2019). "Genome-wide enhancer annotations differ significantly in genomic distribution, evolution, and function". In: *BMC Genomics* 20.1, p. 511. ISSN: 1471-2164. DOI: 10.1186/s12864-019-5779-x.

Berezikov, E., V. Guryev, R. H. A. Plasterk, and E. Cuppen (2004). "CONREAL: Conserved Regulatory Elements Anchored Alignment Algorithm for Identification of Transcription Factor Binding Sites by Phylogenetic Footprinting". In: *Genome Res* 14.1, pp. 170–8. ISSN: 1088-9051 (Print). DOI: 10.1101/gr.1642804.

Bernstein, B. E., T. S. Mikkelsen, X. Xie, M. Kamal, et al. (2006). "A bivalent chromatin structure marks key developmental genes in embryonic stem cells". In: *Cell* 125.2, pp. 315–26. ISSN: 0092-8674 (Print) 0092-8674. DOI: 10.1016/j.cell.2006.02.041.

Bernstein, B. E., J. A. Stamatoyannopoulos, J. F. Costello, B. Ren, et al. (2010). "The NIH Roadmap Epigenomics Mapping Consortium". In: *Nat Biotechnol* 28.10, pp. 1045–8. ISSN: 1087-0156 (Print). DOI: 10.1038/nbt1010-1045.

Berry, F. B., Y. Miura, K. Mihara, P. Kaspar, et al. (2001). "Positive and negative regulation of myogenic differentiation of C2C12 cells by isoforms of the multiple homeodomain zinc finger transcription factor ATBF1". In: *J Biol Chem* 276.27, pp. 25057–65. ISSN: 0021-9258 (Print) 0021-9258. DOI: 10.1074/jbc.M010378200.

Bhagwat, A. S., J. S. Roe, B. Y. L. Mok, A. F. Hohmann, et al. (2016). "BET Bromodomain Inhibition Releases the Mediator Complex from Select cis-Regulatory Elements". In: *Cell Rep* 15.3, pp. 519–530. DOI: 10.1016/j.celrep.2016.03.054.

Bhatia, S. and D. A. Kleinjan (2014). "Disruption of long-range gene regulation in human genetic disease". In: *Human Genetics* 133.7, pp. 815–845. ISSN: 0340-6717. DOI: 10.1007/s00439-014-1424-6.

Billot, K., J. Soeur, F. Chereau, I. Arrouss, et al. (2011). "Deregulation of Aiolos expression in chronic lymphocytic leukemia is associated with epigenetic modifications". In: *Blood* 117.6, pp. 1917–27. ISSN: 0006-4971. DOI: 10.1182/blood-2010-09-307140.

Birney, E., J. A. Stamatoyannopoulos, A. Dutta, R. Guigo, et al. (2007). "Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project". In: *Nature* 447.7146, pp. 799–816. ISSN: 0028-0836. DOI: 10.1038/nature05874.

Blackwood, E. M. and J. T. Kadonaga (1998). "Going the distance: a current view of enhancer action". In: *Science* 281.5373, pp. 60–3. ISSN: 0036-8075 (Print) 0036-8075.

Blake, J. A., J. T. Eppig, J. A. Kadin, J. E. Richardson, et al. (2017). "Mouse Genome Database (MGD)-2017: community knowledge resource for the laboratory mouse". In: *Nucleic Acids Res* 45.D1, pp. D723–d729. ISSN: 0305-1048. DOI: 10.1093/nar/gkw1040.

Blattler, A., L. Yao, Y. Wang, Z. Ye, et al. (2013). "ZBTB33 binds unmethylated regions of the genome associated with actively expressed genes". In: *Epigenetics Chromatin* 6, p. 13. DOI: 10.1186/1756-8935-6-13.

Blow, M. J., D. J. McCulley, Z. Li, T. Zhang, et al. (2010). "ChIP-Seq identification of weakly conserved heart enhancers". In: *Nat Genet* 42.9, pp. 806–10. ISSN: 1061-4036. DOI: 10.1038/ng.650.

Bogu, G. K., P. Vizan, L. W. Stanton, M. Beato, et al. (2015). "Chromatin and RNA Maps Reveal Regulatory Long Noncoding RNAs in Mouse". In: *Mol Cell Biol* 36.5, pp. 809–19. ISSN: 0270-7306. DOI: 10.1128/mcb.00955-15.

Bonn, S., R. P. Zinzen, C. Girardot, E. H. Gustafson, et al. (2012). "Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development". In: *Nat Genet* 44.2, pp. 148–56. ISSN: 1061-4036. DOI: 10.1038/ng.1064.

Boyle, A. P., S. Davis, H. P. Shulha, P. Meltzer, et al. (2008). "High-resolution mapping and characterization of open chromatin across the genome". In: *Cell* 132.2, pp. 311–22. ISSN: 0092-8674. DOI: 10.1016/j.cell.2007.12.014.

Boyle, E. A., Y. I. Li, and J. K. Pritchard (2017). "An Expanded View of Complex Traits: From Polygenic to Omnigenic". In: *Cell* 169.7, pp. 1177–1186. ISSN: 0092-8674. DOI: 10.1016/j.cell.2017.05.038.

Brown, J. D., C. Y. Lin, Q. Duan, G. Griffin, et al. (2014). "NF-kappaB directs dynamic super enhancer formation in inflammation and atherogenesis". In: *Mol Cell* 56.2, pp. 219–231. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2014.08.024.

Brown, S. D. and M. W. Moore (2012a). "The International Mouse Phenotyping Consortium: past and future perspectives on mouse phenotyping". In: *Mamm Genome* 23.9-10, pp. 632–40. ISSN: 0938-8990 (Print) 0938-8990. DOI: 10.1007/s00335-012-9427-x.

Brown, S. D. and P. M. Nolan (1998). "Mouse mutagenesis-systematic studies of mammalian gene function". In: *Hum Mol Genet* 7.10, pp. 1627–33. ISSN: 0964-6906 (Print) 0964-6906.

Brown, S. D. M. and M. W. Moore (2012b). "Towards an encyclopaedia of mammalian gene function: the International Mouse Phenotyping Consortium". In: *Disease Models and Mechanisms* 5.3, pp. 289–292.

Brown, T. M., C. S. Colwell, J. A. Waschek, and H. D. Piggins (2007). "Disrupted Neuronal Activity Rhythms in the Suprachiasmatic Nuclei of Vasoactive Intestinal

Polypeptide-Deficient Mice". In: *J Neurophysiol* 97.3, pp. 2553–8. ISSN: 0022-3077 (Print). DOI: 10.1152/jn.01206.2006.

Brown, T. M., A. T. Hughes, and H. D. Piggins (2005). "Gastrin-releasing peptide promotes suprachiasmatic nuclei cellular rhythmicity in the absence of vasoactive intestinal polypeptide-VPAC2 receptor signaling". In: *J Neurosci* 25.48, pp. 11155–64. ISSN: 0270-6474. DOI: 10.1523/jneurosci.3821-05.2005.

Buenrostro, J. D., B. Wu, H. Y. Chang, and W. J. Greenleaf (2015). "ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide". In: *Curr Protoc Mol Biol* 109, pp. 21.29.1–9. ISSN: 1934-3639 (Print) 1934-3647. DOI: 10.1002/0471142727.mb2129s109.

Buil, A., A. A. Brown, T. Lappalainen, A. Vinuela, et al. (2015). "Gene-gene and gene-environment interactions detected by transcriptome sequence analysis in twins". In: *Nat Genet* 47.1, pp. 88–91. ISSN: 1061-4036. DOI: 10.1038/ng.3162.

Butler, J. E. and J. T. Kadonaga (2001). "Enhancer-promoter specificity mediated by DPE or TATA core promoter motifs". In: *Genes Dev* 15.19, pp. 2515–9. ISSN: 0890-9369 (Print) 0890-9369. DOI: 10.1101/gad.924301.

— (2002). "The RNA polymerase II core promoter: a key component in the regulation of gene expression". In: *Genes Dev* 16.20, pp. 2583–92. ISSN: 0890-9369 (Print) 0890-9369. DOI: 10.1101/gad.1026202.

Cai, X., L. Hou, N. Su, H. Hu, et al. (2010). "Systematic identification of conserved motif modules in the human genome". In: *BMC Genomics* 11, p. 567. DOI: 10.1186/1471-2164-11-567.

Cannavo, E., P. Khoueiry, D. A. Garfield, P. Geeleher, et al. (2016). "Shadow Enhancers Are Pervasive Features of Developmental Regulatory Networks". In: *Curr Biol* 26.1, pp. 38–51. ISSN: 0960-9822. DOI: 10.1016/j.cub.2015.11.034.

Cannon, B. and J. Nedergaard (2004). "Brown adipose tissue: function and physiological significance". In: *Physiol Rev* 84.1, pp. 277–359. ISSN: 0031-9333 (Print) 0031-9333. DOI: 10.1152/physrev.00015.2003.

Canver, M. C., E. C. Smith, F. Sher, L. Pinello, et al. (2015). "BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis". In: *Nature* 527, p. 192. DOI: 10.1038/nature15521https://www.nature.com/articles/nature15521#supplementary-information.

Carter, D., L. Chakalova, C. S. Osborne, Y. F. Dai, et al. (2002). "Long-range chromatin regulatory interactions in vivo". In: *Nat Genet* 32.4, pp. 623–6. ISSN: 1061-4036 (Print) 1061-4036. DOI: 10.1038/ng1051.

Cavalli, G., M. Hayashi, Y. Jin, D. Yorgov, et al. (2016). "MHC class II super-enhancer increases surface expression of HLA-DR and HLA-DQ and affects cytokine production in autoimmune vitiligo". In: *Proc Natl Acad Sci U S A* 113.5, pp. 1363–8. ISSN: 0027-8424. DOI: 10.1073/pnas.1523482113.

Chapuy, B., M. R. McKeown, C. Y. Lin, S. Monti, et al. (2013). "Discovery and characterization of super-enhancer-associated dependencies in diffuse large B cell lymphoma". In: *Cancer Cell* 24.6, pp. 777–90. ISSN: 1535-6108. DOI: 10.1016/j.ccr.2013.11.003.

Chen, J., E. E. Bardes, B. J. Aronow, and A. G. Jegga (2009). "ToppGene Suite for gene list enrichment analysis and candidate gene prioritization". In: *Nucleic Acids Res* 37.Web Server issue, W305–11. ISSN: 0305-1048. DOI: 10.1093/nar/gkp427.

Chen, Y., B. Yao, Z. Zhu, Y. Yi, et al. (2004). "A constitutive super-enhancer: homologous region 3 of Bombyx mori nucleopolyhedrovirus". In: *Biochem Biophys Res Commun* 318.4, pp. 1039–44. ISSN: 0006-291X (Print) 0006-291x. DOI: 10.1016/j.bbrc.2004.04.136.

Cheng, Y., Z. Ma, B.-H. Kim, W. Wu, et al. (2014). "Principles of regulatory information conservation between mouse and human". In: *Nature* 515, p. 371. DOI: 10.1038/nature13985https://www.nature.com/articles/nature13985#supplementary-information.

Chepelev, I., G. Wei, D. Wangsa, Q. Tang, et al. (2012). "Characterization of genome-wide enhancer-promoter interactions reveals co-expression of interacting genes and modes of higher order chromatin organization". In: *Cell Res* 22.3, pp. 490–503. ISSN: 1001-0602. DOI: 10.1038/cr.2012.15.

Chipumuro, E., E. Marco, C. L. Christensen, N. Kwiatkowski, et al. (2014). "CDK7 inhibition suppresses super-enhancer-linked oncogenic transcription in MYCN-driven cancer". In: *Cell* 159.5, pp. 1126–1139. ISSN: 0092-8674. DOI: 10.1016/j.cell.2014.10.024.

Christensen, C. L., N. Kwiatkowski, B. J. Abraham, J. Carretero, et al. (2014). "Targeting transcriptional addictions in small cell lung cancer with a covalent CDK7 inhibitor". In: *Cancer Cell* 26.6, pp. 909–922. ISSN: 1535-6108. DOI: 10.1016/j.ccell.2014.10.019.

Cirillo, L. A., F. R. Lin, I. Cuesta, D. Friedman, et al. (2002). "Opening of compacted chromatin by early developmental transcription factors HNF3 (FoxA) and GATA-4". In: *Mol Cell* 9.2, pp. 279–89. ISSN: 1097-2765 (Print) 1097-2765.

Civelek, M. and A. J. Lusis (2011). "Conducting the metabolic syndrome orchestra". In: *Nat Genet* 43.6, pp. 506–8. ISSN: 1061-4036. DOI: 10.1038/ng.842.

Claussnitzer, M., S. N. Dankel, B. Klocke, H. Grallert, et al. (2014). "Leveraging cross-species transcription factor binding site patterns: from diabetes risk loci to disease mechanisms". In: *Cell* 156.1-2, pp. 343–58. ISSN: 0092-8674. DOI: 10.1016/j.cell.2013.10.058.

Comings, D. E. (1972). "The structure and function of chromatin". In: *Adv Hum Genet* 3, pp. 237–431. ISSN: 0065-275X (Print) 0065-275x.

Connelly, S. and J. L. Manley (1988). "A functional mRNA polyadenylation signal is required for transcription termination by RNA polymerase II". In: *Genes Dev* 2.4, pp. 440–52. ISSN: 0890-9369 (Print) 0890-9369.

Cooper, G. M. and C. D. Brown (2008). "Qualifying the relationship between sequence conservation and molecular function". In: *Genome Res* 18.2, pp. 201–5. ISSN: 1088-9051 (Print) 1088-9051. DOI: 10.1101/gr.7205808.

Cotney, J., J. Leng, S. Oh, L. E. DeMare, et al. (2012). "Chromatin state signatures associated with tissue-specific gene expression and enhancer activity in the embryonic limb". In: *Genome Res* 22.6, pp. 1069–80. ISSN: 1088-9051. DOI: 10.1101/gr.129817.111.

Cotney, J., J. Leng, J. Yin, S. K. Reilly, et al. (2013). "The evolution of lineage-specific regulatory activities in the human embryonic limb". In: *Cell* 154.1, pp. 185–96. ISSN: 0092-8674. DOI: 10.1016/j.cell.2013.05.056.

Cousminer, D. L., D. J. Berry, N. J. Timpson, W. Ang, et al. (2013). "Genome-wide association and longitudinal analyses reveal genetic loci linking pubertal height growth,

pubertal timing and childhood adiposity". In: *Hum Mol Genet* 22.13, pp. 2735–47. ISSN: 0964-6906. DOI: 10.1093/hmg/ddt104.

Creyghton, M. P., A. W. Cheng, G. G. Welstead, T. Kooistra, et al. (2010). "Histone H3K27ac separates active from poised enhancers and predicts developmental state". In: *Proc Natl Acad Sci U S A* 107.50, pp. 21931–6. ISSN: 0027-8424. DOI: 10.1073/pnas.1016071107.

Crick, F. (1970). "Central dogma of molecular biology". In: *Nature* 227.5258, pp. 561–3. ISSN: 0028-0836 (Print) 0028-0836.

Crick, F. H., L. Barnett, S. Brenner, and R. J. Watts-Tobin (1961). "General nature of the genetic code for proteins". In: *Nature* 192, pp. 1227–32. ISSN: 0028-0836 (Print) 0028-0836.

Cunningham, T. J., J. J. Lancman, M. Berenguer, P. D. S. Dong, et al. (2018). "Genomic Knockout of Two Presumed Forelimb Tbx5 Enhancers Reveals They Are Nonessential for Limb Development". In: *Cell Rep* 23.11, pp. 3146–3151. DOI: 10.1016/j.celrep.2018.05.052.

Dang, D. T., J. Pevsner, and V. W. Yang (2000). "The biology of the mammalian Kruppel-like family of transcription factors". In: *Int J Biochem Cell Biol* 32.11-12, pp. 1103–21. ISSN: 1357-2725 (Print) 1357-2725.

Darnell J. E., J. (1982). "Variety in the level of gene control in eukaryotic cells". In: *Nature* 297.5865, pp. 365–71. ISSN: 0028-0836 (Print) 0028-0836.

Das, P. M., K. Ramachandran, J. vanWert, and R. Singal (2004). "Chromatin immunoprecipitation assay". In: *Biotechniques* 37.6, pp. 961–9. ISSN: 0736-6205 (Print) 0736-6205. DOI: 10.2144/04376rv01.

Davidson, I., C. Fromental, P. Augereau, A. Wildeman, et al. (1986). "Cell-type specific protein binding to the enhancer of simian virus 40 in nuclear extracts". In: *Nature* 323.6088, pp. 544–8. ISSN: 0028-0836 (Print) 0028-0836. DOI: 10.1038/323544a0.

Dawson, M. A., E. J. Gudgin, S. J. Horton, G. Giotopoulos, et al. (2014). "Recurrent mutations, including NPM1c, activate a BRD4-dependent core transcriptional program in acute myeloid leukemia". In: *Leukemia* 28.2, pp. 311–20. ISSN: 0887-6924. DOI: 10.1038/leu.2013.338.

De Las Rivas, J. and C. Fontanillo (2010). "Protein-protein interactions essentials: key concepts to building and analyzing interactome networks". In: *PLoS Comput Biol* 6.6, e1000807. ISSN: 1553-734x. DOI: 10.1371/journal.pcbi.1000807.

Defossez, P. A., K. F. Kelly, G. J. Filion, R. Perez-Torrado, et al. (2005). "The human enhancer blocker CTC-binding factor interacts with the transcription factor Kaiso". In: *J Biol Chem* 280.52, pp. 43017–23. ISSN: 0021-9258 (Print) 0021-9258. DOI: 10.1074/jbc.M510802200.

Dekker, J., M. A. Marti-Renom, and L. A. Mirny (2013). "Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data". In: *Nat Rev Genet* 14.6, pp. 390–403. ISSN: 1471-0056 (Print) 1471-0056. DOI: 10.1038/nrg3454.

Delabesse, E., S. Ogilvy, M. A. Chapman, S. G. Piltz, et al. (2005). "Transcriptional regulation of the SCL locus: identification of an enhancer that targets the primitive erythroid lineage in vivo". In: *Mol Cell Biol* 25.12, pp. 5215–25. ISSN: 0270-7306 (Print) 0270-7306. DOI: 10.1128/mcb.25.12.5215-5225.2005.

Diao, Y., B. Li, Z. Meng, I. Jung, et al. (2016). "A new class of temporarily phenotypic enhancers identified by CRISPR/Cas9-mediated genetic screening". In: *Genome Res* 26.3, pp. 397–405. ISSN: 1088-9051 (Print). DOI: 10.1101/gr.197152.115.

Dickel, D. E., A. R. Ypsilanti, R. Pla, Y. Zhu, et al. (2018a). "Ultraconserved Enhancers Are Required for Normal Development". In: *Cell* 172.3, 491–499.e15. ISSN: 0092-8674. DOI: 10.1016/j.cell.2017.12.017.

— (2018b). "Ultraconserved Enhancers Are Required for Normal Development". In: *Cell* 172.3, 491–499.e15. ISSN: 0092-8674. DOI: 10.1016/j.cell.2017.12.017.

Dickinson, M. E., A. M. Flenniken, X. Ji, L. Teboul, et al. (2016). "High-throughput discovery of novel developmental phenotypes". In: *Nature* 537.7621, pp. 508–514. ISSN: 0028-0836. DOI: 10.1038/nature19356.

Diéguez-Hurtado, R., K. Kato, B. D. Giaimo, M. Nieminen-Kelhä, et al. (2019). "Loss of the transcription factor RBPJ induces disease-promoting properties in brain pericytes". In: *Nature Communications* 10.1, p. 2817. ISSN: 2041-1723. DOI: 10.1038/s41467-019-10643-w.

Dittrich, M. T., G. W. Klau, A. Rosenwald, T. Dandekar, et al. (2008). "Identifying functional modules in protein-protein interaction networks: an integrated exact approach". In: *Bioinformatics* 24.13, pp. i223–31. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btn161.

Dixon, J. R., S. Selvaraj, F. Yue, A. Kim, et al. (2012). "Topological domains in mammalian genomes identified by analysis of chromatin interactions". In: *Nature* 485.7398, pp. 376–80. ISSN: 0028-0836. DOI: 10.1038/nature11082.

Dong, X. C., K. D. Copps, S. Guo, Y. Li, et al. (2008). "Inactivation of hepatic Foxo1 by insulin signaling is required for adaptive nutrient homeostasis and endocrine growth regulation". In: *Cell Metab* 8.1, pp. 65–76. ISSN: 1550-4131. DOI: 10.1016/j.cmet.2008.06.006.

Dooley, J., J. E. Garcia-Perez, J. Sreenivasan, S. M. Schlenner, et al. (2016). "The microRNA-29 Family Dictates the Balance Between Homeostatic and Pathological Glucose Handling in Diabetes and Obesity". In: *Diabetes* 65.1, pp. 53–61. DOI: 10.2337/db15-0770.

Dostie, J., T. A. Richmond, R. A. Arnaout, R. R. Selzer, et al. (2006). "Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements". In: *Genome Res* 16.10, pp. 1299–309. ISSN: 1088-9051 (Print) 1088-9051. DOI: 10.1101/gr.5571506.

Douglas, L. N., A. B. McGuire, A. M. Manzardo, and M. G. Butler (2016). "High-resolution chromosome ideogram representation of recognized genes for bipolar disorder". In: *Gene* 586.1, pp. 136–47. ISSN: 0378-1119. DOI: 10.1016/j.gene.2016.04.011.

Dowen, J. M., Z. P. Fan, D. Hnisz, G. Ren, et al. (2014). "Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes". In: *Cell* 159.2, pp. 374–387. ISSN: 0092-8674. DOI: 10.1016/j.cell.2014.09.030.

Drier, Y., M. J. Cotton, K. E. Williamson, S. M. Gillespie, et al. (2016). "An oncogenic MYB feedback loop drives alternate cell fates in adenoid cystic carcinoma". In: *Nat Genet* 48.3, pp. 265–72. ISSN: 1061-4036. DOI: 10.1038/ng.3502.

Dupuis, J., C. Langenberg, I. Prokopenko, R. Saxena, et al. (2010). "New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk". In: *Nat Genet* 42.2, pp. 105–16. ISSN: 1061-4036. DOI: 10.1038/ng.520.

Eguchi, J., Q. W. Yan, D. E. Schones, M. Kamal, et al. (2008). "Interferon-regulatory factors (IRFs) are transcriptional regulators of adipogenesis". In: *Cell Metab* 7.1, pp. 86–94. ISSN: 1550-4131 (Print). DOI: 10.1016/j.cmet.2007.11.002.

Ellard, S. and K. Colclough (2006). "Mutations in the genes encoding the transcription factors hepatocyte nuclear factor 1 alpha (HNF1A) and 4 alpha (HNF4A) in maturity-onset diabetes of the young". In: *Hum Mutat* 27.9, pp. 854–69. ISSN: 1059-7794. DOI: 10.1002/humu.20357.

ENCODE Project Consortium, T. (2012). "An integrated encyclopedia of DNA elements in the human genome". In: *Nature* 489, p. 57. DOI: 10.1038/nature11247https://www.nature.com/articles/nature11247#supplementary-information.

Ernst, J. and M. Kellis (2012). "ChromHMM: automating chromatin-state discovery and characterization". In: *Nat Methods* 9.3, pp. 215–6. ISSN: 1548-7091. DOI: 10.1038/nmeth.1906.

Ernst, J., P. Kheradpour, T. S. Mikkelsen, N. Shoresh, et al. (2011). "Mapping and analysis of chromatin state dynamics in nine human cell types". In: *Nature* 473.7345, pp. 43–49. ISSN: 0028-0836. DOI: http://www.nature.com/nature/journal/v473/n7345/abs/10.1038-nature09906-unlocked.html#supplementary-information.

Fang, Z., K. Hecklau, F. Gross, I. Bachmann, et al. (2015). "Transcription factor co-occupied regions in the murine genome constitute T-helper-cell subtype-specific enhancers". In: *Eur J Immunol* 45.11, pp. 3150–7. ISSN: 0014-2980. DOI: 10.1002/eji.201545713.

FANTOM Consortium, T., t. R. PMI, and CLST (2014). "A promoter-level mammalian expression atlas". In: *Nature* 507, p. 462. DOI: 10.1038/nature13182https://www.nature.com/articles/nature13182#supplementary-information.

Farh, K. K., A. Marson, J. Zhu, M. Kleinewietfeld, et al. (2015). "Genetic and epigenetic fine mapping of causal autoimmune disease variants". In: *Nature* 518.7539, pp. 337–43. ISSN: 0028-0836. DOI: 10.1038/nature13835.

Fisher, W. W., J. J. Li, A. S. Hammonds, J. B. Brown, et al. (2012). "DNA regions bound at low occupancy by transcription factors do not drive patterned reporter gene expression in Drosophila". In: *Proc Natl Acad Sci U S A* 109.52, pp. 21330–5. ISSN: 0027-8424. DOI: 10.1073/pnas.1209589110.

Florez, J. C. (2007). "The new type 2 diabetes gene TCF7L2". In: *Curr Opin Clin Nutr Metab Care* 10.4, pp. 391–6. ISSN: 1363-1950 (Print) 1363-1950. DOI: 10.1097/MCO.0b013e3281e2c9be.

Franceschini, A., D. Szklarczyk, S. Frankild, M. Kuhn, et al. (2013). "STRING v9.1: protein-protein interaction networks, with increased coverage and integration". In: *Nucleic Acids Res* 41.Database issue, pp. D808–15. ISSN: 0305-1048. DOI: 10.1093/nar/gks1094.

Frankel, N., G. K. Davis, D. Vargas, S. Wang, et al. (2010). "Phenotypic robustness conferred by apparently redundant transcriptional enhancers". In: *Nature* 466.7305, pp. 490–3. ISSN: 0028-0836. DOI: 10.1038/nature09158.

Frayling, T. M., M. P. Bulamn, S. Ellard, M. Appleton, et al. (1997). "Mutations in the hepatocyte nuclear factor-1alpha gene are a common cause of maturity-onset

diabetes of the young in the U.K". In: *Diabetes* 46.4, pp. 720–5. ISSN: 0012-1797 (Print) 0012-1797.

Frayling, T. M., M. P. Bulman, M. Appleton, A. T. Hattersley, et al. (1997). "A rapid screening method for hepatocyte nuclear factor 1 alpha frameshift mutations; prevalence in maturity-onset diabetes of the young and late-onset non-insulin dependent diabetes". In: *Hum Genet* 101.3, pp. 351–4. ISSN: 0340-6717 (Print) 0340-6717.

Friedli, M., I. Barde, M. Arcangeli, S. Verp, et al. (2010). "A systematic enhancer screen using lentivector transgenesis identifies conserved and non-conserved functional elements at the Olig1 and Olig2 locus". In: *PLoS One* 5.12, e15741. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0015741.

Friedman, J. M. (2009). "Obesity: Causes and control of excess body fat". In: *Nature* 459.7245, pp. 340–2. ISSN: 0028-0836. DOI: 10.1038/459340a.

Fullwood, M. J., M. H. Liu, Y. F. Pan, J. Liu, et al. (2009). "An oestrogen-receptor-alpha-bound human chromatin interactome". In: *Nature* 462.7269, pp. 58–64. ISSN: 0028-0836. DOI: 10.1038/nature08497.

Gaulton, K. J., T. Nammo, L. Pasquali, J. M. Simon, et al. (2010). "A map of open chromatin in human pancreatic islets". In: *Nat Genet* 42.3, pp. 255–9. ISSN: 1061-4036. DOI: 10.1038/ng.530.

Giraldo, P. and L. Montoliu (2001). "Size matters: use of YACs, BACs and PACs in transgenic animals". In: *Transgenic Res* 10.2, pp. 83–103. ISSN: 0962-8819 (Print) 0962-8819.

Giresi, P. G., J. Kim, R. M. McDaniell, V. R. Iyer, et al. (2007). "FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin". In: *Genome Res* 17.6, pp. 877–85. ISSN: 1088-9051 (Print) 1088-9051. DOI: 10.1101/gr.5533506.

Gjoneska, E., A. R. Pfenning, H. Mathys, G. Quon, et al. (2015). "Conserved epigenomic signals in mice and humans reveal immune basis of Alzheimer's disease". In: *Nature* 518.7539, pp. 365–9. ISSN: 0028-0836. DOI: 10.1038/nature14252.

Goes, F. S., J. McGrath, D. Avramopoulos, P. Wolyniec, et al. (2015). "Genome-wide association study of schizophrenia in Ashkenazi Jews". In: *Am J Med Genet B Neuropsychiatr Genet* 168.8, pp. 649–59. ISSN: 1552-4841. DOI: 10.1002/ajmg.b.32349.

Gondor, A. and R. Ohlsson (2009). "Chromosome crosstalk in three dimensions". In: *Nature* 461.7261, pp. 212–7. ISSN: 0028-0836. DOI: 10.1038/nature08453.

Gonzalez, M. W. and M. G. Kann (2012). "Chapter 4: Protein interactions and disease". In: *PLoS Comput Biol* 8.12, e1002819. ISSN: 1553-734x. DOI: 10.1371/journal.pcbi.1002819.

Gordon, C. T., C. Attanasio, S. Bhatia, S. Benko, et al. (2014). "Identification of novel craniofacial regulatory domains located far upstream of SOX9 and disrupted in Pierre Robin sequence". In: *Hum Mutat* 35.8, pp. 1011–20. ISSN: 1059-7794. DOI: 10.1002/humu.22606.

Gottgens, B., C. Broccardo, M. J. Sanchez, S. Deveaux, et al. (2004). "The scl +18/19 stem cell enhancer is not required for hematopoiesis: identification of a 5' bifunctional hematopoietic-endothelial enhancer bound by Fli-1 and Elf-1". In: *Mol Cell Biol* 24.5, pp. 1870–83. ISSN: 0270-7306 (Print) 0270-7306.

Grant, C. E., T. L. Bailey, and W. S. Noble (2011). "FIMO: scanning for occurrences of a given motif". In: *Bioinformatics* 27.7, pp. 1017–8. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btr064.

Grant, S. F., G. Thorleifsson, I. Reynisdottir, R. Benediktsson, et al. (2006). "Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes". In: *Nat Genet* 38.3, pp. 320–3. ISSN: 1061-4036 (Print) 1061-4036. DOI: 10.1038/ng1732.

Gray, S. and M. Levine (1996). "Transcriptional repression in development". In: *Curr Opin Cell Biol* 8.3, pp. 358–64. ISSN: 0955-0674 (Print) 0955-0674.

Green, R. and H. F. Noller (1997). "Ribosomes and translation". In: *Annu Rev Biochem* 66, pp. 679–716. ISSN: 0066-4154 (Print) 0066-4154. DOI: 10.1146/annurev.biochem.66.1.679.

Groop, L. (2010). "Open chromatin and diabetes risk". In: *Nat Genet* 42.3, pp. 190–2. ISSN: 1061-4036. DOI: 10.1038/ng0310-190.

Groschel, S., M. A. Sanders, R. Hoogenboezem, E. de Wit, et al. (2014). "A single oncogenic enhancer rearrangement causes concomitant EVI1 and GATA2 deregulation in leukemia". In: *Cell* 157.2, pp. 369–381. ISSN: 0092-8674. DOI: 10.1016/j.cell.2014.02.019.

Grosveld, F., G. B. van Assendelft, D. R. Greaves, and G. Kollias (1987). "Position-independent, high-level expression of the human beta-globin gene in transgenic mice". In: *Cell* 51.6, pp. 975–85. ISSN: 0092-8674 (Print) 0092-8674.

Groves, C. J., E. Zeggini, J. Minton, T. M. Frayling, et al. (2006). "Association analysis of 6,736 U.K. subjects provides replication and confirms TCF7L2 as a type 2 diabetes susceptibility gene with a substantial effect on individual risk". In: *Diabetes* 55.9, pp. 2640–4. ISSN: 0012-1797 (Print) 0012-1797. DOI: 10.2337/db06-0355.

Guenther, M. G., S. S. Levine, L. A. Boyer, R. Jaenisch, et al. (2007a). "A chromatin landmark and transcription initiation at most promoters in human cells". In: *Cell* 130.1, pp. 77–88. ISSN: 0092-8674 (Print) 0092-8674. DOI: 10.1016/j.cell.2007.05.042.

— (2007b). "A chromatin landmark and transcription initiation at most promoters in human cells". In: *Cell* 130.1, pp. 77–88. ISSN: 0092-8674 (Print) 0092-8674. DOI: 10.1016/j.cell.2007.05.042.

Hadjur, S., L. M. Williams, N. K. Ryan, B. S. Cobb, et al. (2009). "Cohesins form chromosomal cis-interactions at the developmentally regulated IFNG locus". In: *Nature* 460.7253, pp. 410–3. ISSN: 0028-0836. DOI: 10.1038/nature08079.

Hamdan, F. H. and S. A. Johnsen (2018). "Super enhancers - new analyses and perspectives on the low hanging fruit". In: *Transcription* 9.2, pp. 123–130. ISSN: 2154-1272. DOI: 10.1080/21541264.2017.1372044.

Hamosh, A., A. F. Scott, J. S. Amberger, C. A. Bocchini, et al. (2005). "Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders". In: *Nucleic Acids Res* 33.Database issue, pp. D514–7. ISSN: 0305-1048. DOI: 10.1093/nar/gki033.

Han, J., P. Kraft, H. Nan, Q. Guo, et al. (2008). "A genome-wide association study identifies novel alleles associated with hair color and skin pigmentation". In: *PLoS Genet* 4.5, e1000074. ISSN: 1553-7390. DOI: 10.1371/journal.pgen.1000074.

Hannenhalli, S. and K. H. Kaestner (2009). "The evolution of Fox genes and their role in development and disease". In: *Nat Rev Genet* 10.4, pp. 233–40. ISSN: 1471-0056 (Print). DOI: 10.1038/nrg2523.

Hansen, U. and P. A. Sharp (1983). "Sequences controlling in vitro transcription of SV40 promoters". In: *Embo j* 2.12, pp. 2293–303. ISSN: 0261-4189 (Print) 0261-4189.

Hargreaves, D. C. and G. R. Crabtree (2011). "ATP-dependent chromatin remodeling: genetics, genomics and mechanisms". In: *Cell Res* 21.3, pp. 396–420. ISSN: 1001-0602. DOI: 10.1038/cr.2011.32.

Harp, J. B., D. Franklin, A. A. Vanderpuije, and J. M. Gimble (2001). "Differential expression of signal transducers and activators of transcription during human adipogenesis". In: *Biochem Biophys Res Commun* 281.4, pp. 907–12. ISSN: 0006-291X (Print) 0006-291x. DOI: 10.1006/bbrc.2001.4460.

Hashimoto, H., Z. Wang, G. A. Garry, V. S. Malladi, et al. (2019). "Cardiac Reprogramming Factors Synergistically Activate Genome-wide Cardiogenic Stage-Specific Enhancers". In: *Cell Stem Cell* 25.1, 69–86.e5. ISSN: 1934-5909. DOI: https://doi.org/10.1016/j.stem.2019.03.022.

Hatzis, P. and I. Talianidis (2002). "Dynamics of enhancer-promoter communication during differentiation-induced gene activation". In: *Mol Cell* 10.6, pp. 1467–77. ISSN: 1097-2765 (Print) 1097-2765.

Hay, D., J. R. Hughes, C. Babbs, J. O. J. Davies, et al. (2016). "Genetic dissection of the alpha-globin super-enhancer in vivo". In: *Nat Genet* 48.8, pp. 895–903. ISSN: 1061-4036. DOI: 10.1038/ng.3605.

He, A., L. Zhu, N. Gupta, Y. Chang, et al. (2007). "Overexpression of micro ribonucleic acid 29, highly up-regulated in diabetic rats, leads to insulin resistance in 3T3-L1 adipocytes". In: *Mol Endocrinol* 21.11, pp. 2785–94. ISSN: 0888-8809 (Print) 0888-8809. DOI: 10.1210/me.2007-0167.

He, Q., A. F. Bardet, B. Patton, J. Purvis, et al. (2011). "High conservation of transcription factor binding and evidence for combinatorial regulation across six Drosophila species". In: *Nat Genet* 43.5, pp. 414–20. ISSN: 1061-4036. DOI: 10.1038/ng.808.

Heintzman, N. D., G. C. Hon, R. D. Hawkins, P. Kheradpour, et al. (2009). "Histone Modifications at Human Enhancers Reflect Global Cell Type-Specific Gene Expression". In: *Nature* 459.7243, pp. 108–12. ISSN: 0028-0836 (Print). DOI: 10.1038/nature07829.

Heintzman, N. D., R. K. Stuart, G. Hon, Y. Fu, et al. (2007). "Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome". In: *Nat Genet* 39.3, pp. 311–8. ISSN: 1061-4036 (Print) 1061-4036. DOI: 10.1038/ng1966.

Heinz, S., C. Benner, N. Spann, E. Bertolino, et al. (2010). "Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities". In: *Mol Cell* 38.4, pp. 576–89. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2010.05.004.

Heinz, S., C. E. Romanoski, C. Benner, and C. K. Glass (2015). "The selection and function of cell type-specific enhancers". In: *Nat Rev Mol Cell Biol* 16.3, pp. 144–54. ISSN: 1471-0072. DOI: 10.1038/nrm3949.

Herz, H. M. (2016). "Enhancer deregulation in cancer and other diseases". In: *Bioessays* 38.10, pp. 1003–15. ISSN: 0265-9247. DOI: 10.1002/bies.201600106.

Hesselberth, J. R., X. Chen, Z. Zhang, P. J. Sabo, et al. (2009). "Global mapping of protein-DNA interactions in vivo by digital genomic footprinting". In: *Nat Methods* 6.4, pp. 283–9. ISSN: 1548-7091. DOI: 10.1038/nmeth.1313.

Hindorff, L. A., P. Sethupathy, H. A. Junkins, E. M. Ramos, et al. (2009). "Potential etiologic and functional implications of genome-wide association loci for human diseases and traits". In: *Proc Natl Acad Sci U S A* 106.23, pp. 9362–7. ISSN: 0027-8424. DOI: 10.1073/pnas.0903103106.

Hnisz, D., B. J. Abraham, T. I. Lee, A. Lau, et al. (2013). "Super-enhancers in the control of cell identity and disease". In: *Cell* 155.4, pp. 934–47. ISSN: 0092-8674. DOI: 10.1016/j.cell.2013.09.053.

Hnisz, D., J. Schuijers, C. Y. Lin, A. S. Weintraub, et al. (2015). "Convergence of developmental and oncogenic signaling pathways at transcriptional super-enhancers". In: *Mol Cell* 58.2, pp. 362–70. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2015.02.014.

Hong, J. W., D. A. Hendrix, and M. S. Levine (2008). "Shadow enhancers as a source of evolutionary novelty". In: *Science* 321.5894, p. 1314. ISSN: 0036-8075. DOI: 10.1126/science.1160631.

Hoogenkamp, M., M. Lichtinger, H. Krysinska, C. Lancrin, et al. (2009). "Early chromatin unfolding by RUNX1: a molecular explanation for differential requirements during specification versus maintenance of the hematopoietic gene expression program". In: *Blood* 114.2, pp. 299–309. ISSN: 0006-4971. DOI: 10.1182/blood-2008-11-191890.

Hou, C., R. Dale, and A. Dean (2010). "Cell type specificity of chromatin organization mediated by CTCF and cohesin". In: *Proc Natl Acad Sci U S A* 107.8, pp. 3651–6. ISSN: 0027-8424. DOI: 10.1073/pnas.0912087107.

Hrabe de Angelis, M. H., H. Flaswinkel, H. Fuchs, B. Rathkolb, et al. (2000). "Genome-wide, large-scale production of mutant mice by ENU mutagenesis". In: *Nat Genet* 25.4, pp. 444–7. ISSN: 1061-4036 (Print) 1061-4036. DOI: 10.1038/78146.

Hsieh, C. L., T. Fei, Y. Chen, T. Li, et al. (2014). "Enhancer RNAs participate in androgen receptor-driven looping that selectively enhances gene activation". In: *Proc Natl Acad Sci U S A* 111.20, pp. 7319–24. ISSN: 0027-8424. DOI: 10.1073/pnas.1324151111.

Hsu, J., J. Arand, A. Chaikovsky, N. A. Mooney, et al. (2019). "E2F4 regulates transcriptional activation in mouse embryonic stem cells independently of the RB family". In: *Nature Communications* 10.1, p. 2939. ISSN: 2041-1723. DOI: 10.1038/s41467-019-10901-x.

Huang, J., X. Liu, D. Li, Z. Shao, et al. (2016). "Dynamic Control of Enhancer Repertoires Drives Lineage and Stage-Specific Transcription during Hematopoiesis". In: *Dev Cell* 36.1, pp. 9–23. ISSN: 1534-5807. DOI: 10.1016/j.devcel.2015.12.014.

Huang, Q., T. Whitington, P. Gao, J. F. Lindberg, et al. (2014). "A prostate cancer susceptibility allele at 6q22 increases RFX6 expression by modulating HOXB13 chromatin binding". In: *Nat Genet* 46.2, pp. 126–35. ISSN: 1061-4036. DOI: 10.1038/ng.2862.

Hughes, J. D., P. W. Estep, S. Tavazoie, and G. M. Church (2000). "Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae". In: *J Mol Biol* 296.5, pp. 1205–14. ISSN: 0022-2836 (Print) 0022-2836. DOI: 10.1006/jmbi.2000.3519.

Istaces, N., M. Splittgerber, V. Lima Silva, M. Nguyen, et al. (2019). "EOMES interacts with RUNX3 and BRG1 to promote innate memory cell formation through epigenetic reprogramming". In: *Nature Communications* 10.1, p. 3306. ISSN: 2041-1723. DOI: 10.1038/s41467-019-11233-6.

Iwakawa, R., M. Takenaka, T. Kohno, Y. Shimada, et al. (2013). "Genome-wide identification of genes with amplification and/or fusion in small cell lung cancer". In: *Genes Chromosomes Cancer* 52.9, pp. 802–16. ISSN: 1045-2257. DOI: 10.1002/gcc.22076.

Jain, D., S. Baldi, A. Zabel, T. Straub, et al. (2015). "Active promoters give rise to false positive 'Phantom Peaks' in ChIP-seq experiments". In: *Nucleic Acids Res* 43.14, pp. 6959–68. ISSN: 0305-1048 (Print). DOI: 10.1093/nar/gkv637.

Jia, L., G. Landan, M. Pomerantz, R. Jaschek, et al. (2009). "Functional enhancers at the gene-poor 8q24 cancer-linked locus". In: *PLoS Genet* 5.8, e1000597. ISSN: 1553-7390. DOI: 10.1371/journal.pgen.1000597.

Jiang, Y. Y., D. C. Lin, A. Mayakonda, M. Hazawa, et al. (2017). "Targeting super-enhancer-associated oncogenes in oesophageal squamous cell carcinoma". In: *Gut* 66.8, pp. 1358–1368. ISSN: 0017-5749. DOI: 10.1136/gutjnl-2016-311818.

Jiao, W., Y. Chen, H. Song, D. Li, et al. (2018). "HPSE enhancer RNA promotes cancer progression through driving chromatin looping and regulating hnRNPU/p300/EGR1/HPSE axis". In: *Oncogene* 37.20, pp. 2728–2745. ISSN: 0950-9232. DOI: 10.1038/s41388-018-0128-0.

Jin, F., Y. Li, J. R. Dixon, S. Selvaraj, et al. (2013). "A high-resolution map of the three-dimensional chromatin interactome in human cells". In: *Nature* 503.7475, pp. 290–4. ISSN: 0028-0836. DOI: 10.1038/nature12644.

John, S., P. J. Sabo, R. E. Thurman, M. H. Sung, et al. (2011). "Chromatin accessibility pre-determines glucocorticoid receptor binding patterns". In: *Nat Genet* 43.3, pp. 264–8. ISSN: 1061-4036. DOI: 10.1038/ng.759.

Johnson, D. S., A. Mortazavi, R. M. Myers, and B. Wold (2007). "Genome-wide mapping of in vivo protein-DNA interactions". In: *Science* 316.5830, pp. 1497–502. ISSN: 0036-8075. DOI: 10.1126/science.1141319.

Jolma, A., J. Yan, T. Whitington, J. Toivonen, et al. (2013). "DNA-binding specificities of human transcription factors". In: *Cell* 152.1-2, pp. 327–39. ISSN: 0092-8674. DOI: 10.1016/j.cell.2012.12.009.

Jung, C. G., H. J. Kim, M. Kawaguchi, K. K. Khanna, et al. (2005). "Homeotic factor ATBF1 induces the cell cycle arrest associated with neuronal differentiation". In: *Development* 132.23, pp. 5137–45. ISSN: 0950-1991 (Print) 0950-1991. DOI: 10.1242/dev.02098.

Kadonaga, J. T. (2004). "Regulation of RNA polymerase II transcription by sequence-specific DNA binding factors". In: *Cell* 116.2, pp. 247–57. ISSN: 0092-8674 (Print) 0092-8674.

Kagey, M. H., J. J. Newman, S. Bilodeau, Y. Zhan, et al. (2010). "Mediator and cohesin connect gene expression and chromatin architecture". In: *Nature* 467.7314, pp. 430–5. ISSN: 0028-0836. DOI: 10.1038/nature09380.

Kaplan, T., X. Y. Li, P. J. Sabo, S. Thomas, et al. (2011). "Quantitative models of the mechanisms that control genome-wide patterns of transcription factor binding during early Drosophila development". In: *PLoS Genet* 7.2, e1001290. ISSN: 1553-7390. DOI: 10.1371/journal.pgen.1001290.

Kaspar, P., M. Dvorakova, J. Kralova, P. Pajer, et al. (1999). "Myb-interacting protein, ATBF1, represses transcriptional activity of Myb oncoprotein". In: *J Biol Chem* 274.20, pp. 14422–8. ISSN: 0021-9258 (Print) 0021-9258.

Kellum, R. and P. Schedl (1992). "A group of scs elements function as domain boundaries in an enhancer-blocking assay". In: *Molecular and Cellular Biology* 12.5, p. 2424.

Khan, A. and X. Zhang (2016). "dbSUPER: a database of super-enhancers in mouse and human genome". In: *Nucleic Acids Res* 44.D1, pp. D164–71. ISSN: 0305-1048. DOI: 10.1093/nar/gkv1002.

Khan, A. and X. Zhang (2017). "Integrative analysis reveals genomic and epigenomic signatures of super-enhancers and its constituents". In: *bioRxiv*.

Kharchenko, P. V., A. A. Alekseyenko, Y. B. Schwartz, A. Minoda, et al. (2010). "Comprehensive analysis of the chromatin landscape in Drosophila melanogaster". In: *Nature* 471, p. 480. DOI: 10.1038/nature09725https://www.nature.com/articles/nature09725#supplementary-information.

Kim, K. K., R. S. Adelstein, and S. Kawamoto (2009). "Identification of neuronal nuclei (NeuN) as Fox-3, a new member of the Fox-1 gene family of splicing factors". In: *J Biol Chem* 284.45, pp. 31052–61. ISSN: 0021-9258. DOI: 10.1074/jbc.M109.052969.

Kim, K. K., J. Nam, Y. S. Mukouyama, and S. Kawamoto (2013). "Rbfox3-regulated alternative splicing of Numb promotes neuronal differentiation during development". In: *J Cell Biol* 200.4, pp. 443–58. ISSN: 0021-9525. DOI: 10.1083/jcb.201206146.

Kim, M. K., L. A. Lesoon-Wood, B. D. Weintraub, and J. H. Chung (1996). "A soluble transcription factor, Oct-1, is also found in the insoluble nuclear matrix and possesses silencing activity in its alanine-rich domain". In: *Mol Cell Biol* 16.8, pp. 4366–77. ISSN: 0270-7306 (Print) 0270-7306. DOI: 10.1128/mcb.16.8.4366.

Kim, T. H., L. O. Barrera, M. Zheng, C. Qu, et al. (2005). "A high-resolution map of active promoters in the human genome". In: *Nature* 436, p. 876. DOI: 10.1038/nature03877https://www.nature.com/articles/nature03877#supplementary-information.

Kioussis, D., E. Vanin, T. deLange, R. A. Flavell, et al. (1983). "Beta-globin gene inactivation by DNA translocation in gamma beta-thalassaemia". In: *Nature* 306.5944, pp. 662–6. ISSN: 0028-0836 (Print) 0028-0836.

Kleinjan, D. A. and L. A. Lettice (2008). "Long-range gene control and genetic disease". In: *Adv Genet* 61, pp. 339–88. ISSN: 0065-2660 (Print) 0065-2660. DOI: 10.1016/s0065-2660(07)00013-2.

Kleinjan, D. A., A. Seawright, A. Schedl, R. A. Quinlan, et al. (2001). "Aniridia-associated translocations, DNase hypersensitivity, sequence comparison and transgenic analysis redefine the functional domain of PAX6". In: *Hum Mol Genet* 10.19, pp. 2049–59. ISSN: 0964-6906 (Print) 0964-6906.

Kok, Y. J. de, G. F. Merkx, S. M. van der Maarel, I. Huber, et al. (1995). "A duplication/paracentric inversion associated with familial X-linked deafness (DFN3) suggests the presence of a regulatory element more than 400 kb upstream of the POU3F4 gene". In: *Hum Mol Genet* 4.11, pp. 2145–50. ISSN: 0964-6906 (Print) 0964-6906.

Kok, Y. J. de, E. R. Vossenaar, C. W. Cremers, N. Dahl, et al. (1996). "Identification of a hot spot for microdeletions in patients with X-linked deafness type 3 (DFN3)

900 kb proximal to the DFN3 gene POU3F4". In: *Hum Mol Genet* 5.9, pp. 1229–35. ISSN: 0964-6906 (Print) 0964-6906.

Kollias, G., J. Hurst, E. deBoer, and F. Grosveld (1987). "The human beta-globin gene contains a downstream developmental specific enhancer". In: *Nucleic Acids Res* 15.14, pp. 5739–47. ISSN: 0305-1048 (Print) 0305-1048.

Kong, A., V. Steinthorsdottir, G. Masson, G. Thorleifsson, et al. (2009). "Parental origin of sequence variants associated with complex diseases". In: *Nature* 462.7275, pp. 868–74. ISSN: 0028-0836. DOI: 10.1038/nature08625.

Korkmaz, G., R. Lopes, A. P. Ugalde, E. Nevedomskaya, et al. (2016). "Functional genetic screens for enhancer elements in the human genome using CRISPR-Cas9". In: *Nature Biotechnology* 34, p. 192. DOI: 10.1038/nbt.3450https://www.nature.com/articles/nbt.3450#supplementary-information.

Kornberg, R. D. (2007). "The molecular basis of eukaryotic transcription". In: *Proc Natl Acad Sci U S A* 104.32, pp. 12955–61. ISSN: 0027-8424 (Print) 0027-8424. DOI: 10.1073/pnas.0704138104.

Koscielny, G., P. An, D. Carvalho-Silva, J. A. Cham, et al. (2017). "Open Targets: a platform for therapeutic target identification and validation". In: *Nucleic Acids Res* 45.Database issue, pp. D985–94. ISSN: 0305-1048 (Print). DOI: 10.1093/nar/gkw1055.

Kothary, R., S. Clapoff, S. Darling, M. D. Perry, et al. (1989). "Inducible expression of an hsp68-lacZ hybrid gene in transgenic mice". In: *Development* 105.4, pp. 707–14. ISSN: 0950-1991 (Print) 0950-1991.

Kryuchkova-Mostacci, N. and M. Robinson-Rechavi (2017). "A benchmark of gene expression tissue-specificity metrics". In: *Briefings in Bioinformatics* 18.2, pp. 205–214. ISSN: 1467-5463 1477-4054. DOI: 10.1093/bib/bbw008.

Kuhn, M. (2008). "Building Predictive Models in R Using the caret Package". In: *2008* 28.5, p. 26. ISSN: 1548-7660. DOI: 10.18637/jss.v028.i05.

Kulakovskiy, I. V., I. E. Vorontsov, I. S. Yevshin, R. N. Sharipov, et al. (2018). "HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis". In: *Nucleic Acids Res* 46.D1, pp. D252–d259. ISSN: 0305-1048. DOI: 10.1093/nar/gkx1106.

Kulakovskiy, I. V., I. E. Vorontsov, I. S. Yevshin, A. V. Soboleva, et al. (2016). "HOCOMOCO: expansion and enhancement of the collection of transcription factor binding sites models". In: *Nucleic Acids Res* 44.D1, pp. D116–25. ISSN: 0305-1048. DOI: 10.1093/nar/gkv1249.

Kulakovskiy, I. V., Y. A. Medvedeva, U. Schaefer, A. S. Kasianov, et al. (2013). "HOCOMOCO: a comprehensive collection of human transcription factor binding sites models". In: *Nucleic Acids Research* 41.Database issue, pp. D195–D202. ISSN: 0305-1048 1362-4962. DOI: 10.1093/nar/gks1089.

Kumari, M., X. Wang, L. Lantier, A. Lyubetskaya, et al. (2016). "IRF3 promotes adipose inflammation and insulin resistance and represses browning". In: *J Clin Invest* 126.8, pp. 2839–54. ISSN: 0021-9738 (Print). DOI: 10.1172/jci86080.

Kurbatova, N., J. C. Mason, H. Morgan, T. F. Meehan, et al. (2015). "PhenStat: A Tool Kit for Standardized Analysis of High Throughput Phenotypic Data". In: *PLoS One* 10.7, e0131274. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0131274.

Kuska, B. (1998). "Should Scientists Scrap the Notion of Junk DNA?" In: *JNCI: Journal of the National Cancer Institute* 90.14, pp. 1032–1033. ISSN: 0027-8874. DOI: 10.1093/jnci/90.14.1032.

Kvon, E. Z., T. Kazmar, G. Stampfel, J. O. Yanez-Cuna, et al. (2014). "Genome-scale functional characterization of Drosophila developmental enhancers in vivo". In: *Nature* advance online publication. ISSN: 1476-4687. DOI: 10.1038/nature13395http://www.nature.com/nature/journal/vaop/ncurrent/abs/nature13395.html#supplementary-information.

Kwasnieski, J. C., I. Mogno, C. A. Myers, J. C. Corbo, et al. (2012). "Complex effects of nucleotide variants in a mammalian cis-regulatory element". In: *Proceedings of the National Academy of Sciences* 109.47, p. 19498. DOI: 10.1073/pnas.1210678109.

Kwiatkowski, N., T. Zhang, P. B. Rahl, B. J. Abraham, et al. (2014). "Targeting transcription regulation in cancer with a covalent CDK7 inhibitor". In: *Nature* 511.7511, pp. 616–20. ISSN: 0028-0836. DOI: 10.1038/nature13393.

Laat, W. de and F. Grosveld (2003). "Spatial organization of gene expression: the active chromatin hub". In: *Chromosome Res* 11.5, pp. 447–59. ISSN: 0967-3849 (Print) 0967-3849.

Laat, W. de and D. Duboule (2013). "Topology of mammalian developmental enhancers and their regulatory landscapes". In: *Nature* 502.7472, pp. 499–506. ISSN: 0028-0836. DOI: 10.1038/nature12753.

Lage, K., N. T. Hansen, E. O. Karlberg, A. C. Eklund, et al. (2008). "A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes". In: *Proc Natl Acad Sci U S A* 105.52, pp. 20870–5. ISSN: 0027-8424. DOI: 10.1073/pnas.0810772105.

Landrum, M. J., J. M. Lee, G. R. Riley, W. Jang, et al. (2014). "ClinVar: public archive of relationships among sequence variation and human phenotype". In: *Nucleic Acids Res* 42.Database issue, pp. D980–5. ISSN: 0305-1048 (Print) 0305-1048. DOI: 10.1093/nar/gkt1113.

Lawrence, M. S., P. Stojanov, C. H. Mermel, J. T. Robinson, et al. (2014). "Discovery and saturation analysis of cancer genes across 21 tumour types". In: *Nature* 505.7484, pp. 495–501. ISSN: 0028-0836. DOI: 10.1038/nature12912.

Le Gras, S., C. Keime, A. Anthony, C. Lotz, et al. (2017). "Altered enhancer transcription underlies Huntington's disease striatal transcriptional signature". In: *Sci Rep* 7, p. 42875. ISSN: 2045-2322. DOI: 10.1038/srep42875.

Lee, I. T., A. S. Chang, M. Manandhar, Y. Shan, et al. (2015). "Neuromedin s-producing neurons act as essential pacemakers in the suprachiasmatic nucleus to couple clock neurons and dictate circadian rhythms". In: *Neuron* 85.5, pp. 1086–102. ISSN: 0896-6273. DOI: 10.1016/j.neuron.2015.02.006.

Lee, W., A. Haslinger, M. Karin, and R. Tjian (1987). "Activation of transcription by two factors that bind promoter and enhancer sequences of the human metallothionein gene and SV40". In: *Nature* 325.6102, pp. 368–72. ISSN: 0028-0836 (Print) 0028-0836. DOI: 10.1038/325368a0.

Lein, E. S., M. J. Hawrylycz, N. Ao, M. Ayres, et al. (2007). "Genome-wide atlas of gene expression in the adult mouse brain". In: *Nature* 445.7124, pp. 168–76. ISSN: 0028-0836. DOI: 10.1038/nature05453.

Lenhard, B., A. Sandelin, and P. Carninci (2012). "Metazoan promoters: emerging characteristics and insights into transcriptional regulation". In: *Nat Rev Genet* 13.4, pp. 233–45. ISSN: 1471-0056. DOI: 10.1038/nrg3163.

Lettice, L. A., T. Horikoshi, S. J. Heaney, M. J. van Baren, et al. (2002). "Disruption of a long-range cis-acting regulator for Shh causes preaxial polydactyly". In: *Proc Natl Acad Sci U S A* 99.11, pp. 7548–53. ISSN: 0027-8424 (Print) 0027-8424. DOI: 10.1073/pnas.112212199.

Levings, P. P. and J. Bungert (2002). "The human beta-globin locus control region". In: *Eur J Biochem* 269.6, pp. 1589–99. ISSN: 0014-2956 (Print) 0014-2956.

Li, G., X. Ruan, R. K. Auerbach, K. S. Sandhu, et al. (2012). "Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation". In: *Cell* 148.1-2, pp. 84–98. ISSN: 0092-8674. DOI: 10.1016/j.cell.2011.12.014.

Li, J. D., W. P. Hu, L. Boehmer, M. Y. Cheng, et al. (2006). "Attenuated circadian rhythms in mice lacking the prokineticin 2 gene". In: *J Neurosci* 26.45, pp. 11615–23. ISSN: 0270-6474. DOI: 10.1523/jneurosci.3679-06.2006.

Li, Q., K. R. Peterson, X. Fang, and G. Stamatoyannopoulos (2002). "Locus control regions". In: *Blood* 100.9, pp. 3077–86. ISSN: 0006-4971 (Print) 0006-4971. DOI: 10.1182/blood-2002-04-1104.

Li, S., E. Z. Kvon, A. Visel, L. A. Pennacchio, et al. (2019). "Stable enhancers are active in development, and fragile enhancers are associated with evolutionary adaptation". In: *Genome Biology* 20.1, p. 140. ISSN: 1474-760X. DOI: 10.1186/s13059-019-1750-z.

Li, W., D. Notani, Q. Ma, B. Tanasa, et al. (2013). "Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation". In: *Nature* 498.7455, pp. 516–20. ISSN: 0028-0836. DOI: 10.1038/nature12210.

Li, W., L. Yang, Q. He, C. Hu, et al. (2019). "A Homeostatic Arid1a-Dependent Permissive Chromatin State Licenses Hepatocyte Responsiveness to Liver-Injury-Associated YAP Signaling". In: *Cell Stem Cell* 25.1, 54–68.e5. ISSN: 1934-5909. DOI: https://doi.org/10.1016/j.stem.2019.06.008.

Li, X. Y., S. Thomas, P. J. Sabo, M. B. Eisen, et al. (2011). "The role of chromatin accessibility in directing the widespread, overlapping patterns of Drosophila transcription factor binding". In: *Genome Biol* 12.4, R34. ISSN: 1474-7596. DOI: 10.1186/gb-2011-12-4-r34.

Li, Y., C. M. Rivera, H. Ishii, F. Jin, et al. (2014). "CRISPR reveals a distal super-enhancer required for Sox2 expression in mouse embryonic stem cells". In: *PLoS One* 9.12, e114485. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0114485.

Liang, J., H. Zhou, C. Gerdt, M. Tan, et al. (2016). "Epstein-Barr virus super-enhancer eRNAs are essential for MYC oncogene expression and lymphoblast proliferation". In: *Proc Natl Acad Sci U S A* 113.49, pp. 14121–14126. ISSN: 0027-8424. DOI: 10.1073/pnas.1616697113.

Liber, D., R. Domaschenz, P. H. Holmqvist, L. Mazzarella, et al. (2010). "Epigenetic priming of a pre-B cell-specific enhancer through binding of Sox2 and Foxd3 at the ESC stage". In: *Cell Stem Cell* 7.1, pp. 114–26. ISSN: 1875-9777. DOI: 10.1016/j.stem.2010.05.020.

Lieberman-Aiden, E., N. L. van Berkum, L. Williams, M. Imakaev, et al. (2009). "Comprehensive mapping of long-range interactions reveals folding principles of the human genome". In: *Science* 326.5950, pp. 289–93. ISSN: 0036-8075. DOI: 10.1126/science.1181369.

Lilue, J., A. G. Doran, I. T. Fiddes, M. Abrudan, et al. (2018). "Sixteen diverse laboratory mouse reference genomes define strain-specific haplotypes and novel functional loci". In: *Nature Genetics* 50.11, pp. 1574–1583. ISSN: 1546-1718. DOI: 10.1038/s41588-018-0223-8.

Lin, C. Y., J. Loven, P. B. Rahl, R. M. Paranal, et al. (2012). "Transcriptional amplification in tumor cells with elevated c-Myc". In: *Cell* 151.1, pp. 56–67. ISSN: 0092-8674. DOI: 10.1016/j.cell.2012.08.026.

Link, V. M., S. H. Duttke, H. B. Chun, I. R. Holtman, et al. (2018). "Analysis of Genetically Diverse Macrophages Reveals Local and Domain-wide Mechanisms that Control Transcription Factor Binding and Function". In: *Cell* 173.7, 1796–1809.e17. ISSN: 0092-8674. DOI: 10.1016/j.cell.2018.04.018.

Liu, C. F. and V. Lefebvre (2015). "The transcription factors SOX9 and SOX5/SOX6 cooperate genome-wide through super-enhancers to drive chondrogenesis". In: *Nucleic Acids Res* 43.17, pp. 8183–203. ISSN: 0305-1048. DOI: 10.1093/nar/gkv688.

Liu, J. K., C. M. DiPersio, and K. S. Zaret (1991). "Extracellular signals that regulate liver transcription factors during hepatic differentiation in vitro". In: *Mol Cell Biol* 11.2, pp. 773–84. ISSN: 0270-7306 (Print) 0270-7306.

Liu, W., Q. Ma, K. Wong, W. Li, et al. (2013). "Brd4 and JMJD6-associated anti-pause enhancers in regulation of transcriptional pause release". In: *Cell* 155.7, pp. 1581–1595. ISSN: 0092-8674. DOI: 10.1016/j.cell.2013.10.056.

Loots, G. G., M. Kneissel, H. Keller, M. Baptist, et al. (2005). "Genomic deletion of a long-range bone enhancer misregulates sclerostin in Van Buchem disease". In: *Genome Res* 15.7, pp. 928–35. ISSN: 1088-9051 (Print) 1088-9051. DOI: 10.1101/gr.3437105.

Loven, J., H. A. Hoke, C. Y. Lin, A. Lau, et al. (2013). "Selective inhibition of tumor oncogenes by disruption of super-enhancers". In: *Cell* 153.2, pp. 320–34. ISSN: 0092-8674. DOI: 10.1016/j.cell.2013.03.036.

Lowrey, P. L. and J. S. Takahashi (2004). "MAMMALIAN CIRCADIAN BIOLOGY: ELUCIDATING GENOME-WIDE LEVELS OF TEMPORAL ORGANIZATION". In: *Annu Rev Genomics Hum Genet* 5, pp. 407–41. ISSN: 1527-8204 (Print). DOI: 10.1146/annurev.genom.5.061903.175925.

Luizon, M. R. and N. Ahituv (2015). "Uncovering drug-responsive regulatory elements". In: *Pharmacogenomics* 16.16, pp. 1829–41. ISSN: 1462-2416. DOI: 10.2217/pgs.15.121.

Luscombe, N. M., R. A. Laskowski, and J. M. Thornton (2001). "Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level". In: *Nucleic Acids Res* 29.13, pp. 2860–74. ISSN: 0305-1048.

Lyssenko, V., R. Lupi, P. Marchetti, S. Del Guerra, et al. (2007). "Mechanisms by which common variants in the TCF7L2 gene increase risk of type 2 diabetes". In: *J Clin Invest* 117.8, pp. 2155–63. ISSN: 0021-9738 (Print) 0021-9738. DOI: 10.1172/jci30706.

Machanick, P. and T. L. Bailey (2011). "MEME-ChIP: motif analysis of large DNA datasets". In: *Bioinformatics* 27.12, pp. 1696–7. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btr189.

Makinen, N., M. Mehine, J. Tolvanen, E. Kaasinen, et al. (2011). "MED12, the mediator complex subunit 12 gene, is mutated at high frequency in uterine leiomyomas". In: *Science* 334.6053, pp. 252–5. ISSN: 0036-8075. DOI: 10.1126/science.1208930.

Malumbres, M. (2014). "Cyclin-dependent kinases". In: *Genome Biol* 15.6, p. 122. ISSN: 1474-7596.

Maniatis, T., S. Goodbourn, and J. A. Fischer (1987). "Regulation of inducible and tissue-specific gene expression". In: *Science* 236.4806, pp. 1237–45. ISSN: 0036-8075 (Print) 0036-8075.

Manolio, T. A., F. S. Collins, N. J. Cox, D. B. Goldstein, et al. (2009). "Finding the missing heritability of complex diseases". In: *Nature* 461.7265, pp. 747–53. ISSN: 0028-0836. DOI: 10.1038/nature08494.

Manolio, T. A. (2010). "Genomewide Association Studies and Assessment of the Risk of Disease". In: *New England Journal of Medicine* 363.2, pp. 166–176. ISSN: 0028-4793. DOI: 10.1056/NEJMra0905980.

Mansour, M. R., B. J. Abraham, L. Anders, A. Berezovskaya, et al. (2014). "Oncogene regulation. An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element". In: *Science* 346.6215, pp. 1373–7. ISSN: 0036-8075. DOI: 10.1126/science.1259037.

Markowski, D. N., H. W. Thies, A. Gottlieb, H. Wenk, et al. (2013). "HMGA2 expression in white adipose tissue linking cellular senescence with diabetes". In: *Genes Nutr* 8.5, pp. 449–56. ISSN: 1555-8932 (Print) 1555-8932. DOI: 10.1007/s12263-013-0354-6.

Martin, C., P. C. Burdon, G. Bridger, J. C. Gutierrez-Ramos, et al. (2003). "Chemokines acting via CXCR2 and CXCR4 control the release of neutrophils from the bone marrow and their return following senescence". In: *Immunity* 19.4, pp. 583–93. ISSN: 1074-7613 (Print) 1074-7613.

Massart, J., R. J. Sjögren, L. S. Lundell, J. M. Mudry, et al. (2017). "Altered miR-29 Expression in Type 2 Diabetes Influences Glucose and Lipid Metabolism in Skeletal Muscle". In: *Diabetes* 66.7, pp. 1807–1818. DOI: 10.2337/db17-0141.

Maston, G. A., S. K. Evans, and M. R. Green (2006). "Transcriptional regulatory elements in the human genome". In: *Annu Rev Genomics Hum Genet* 7, pp. 29–59. ISSN: 1527-8204 (Print) 1527-8204. DOI: 10.1146/annurev.genom.7.080505.115623.

Masuya, H., M. Inoue, Y. Wada, A. Shimizu, et al. (2005). "Implementation of the modified-SHIRPA protocol for screening of dominant phenotypes in a large-scale ENU mutagenesis program". In: *Mamm Genome* 16.11, pp. 829–37. ISSN: 0938-8990 (Print) 0938-8990. DOI: 10.1007/s00335-005-2430-8.

Mathelier, A., X. Zhao, A. W. Zhang, F. Parcy, et al. (2014). "JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles". In: *Nucleic Acids Res* 42.Database issue, pp. D142–7. ISSN: 0305-1048. DOI: 10.1093/nar/gkt997.

Maurano, M. T., R. Humbert, E. Rynes, R. E. Thurman, et al. (2012). "Systematic localization of common disease-associated variation in regulatory DNA". In: *Science* 337.6099, pp. 1190–5. ISSN: 0036-8075. DOI: 10.1126/science.1222794.

May, D., M. J. Blow, T. Kaplan, D. J. McCulley, et al. (2011). "Large-scale discovery of enhancers from human heart tissue". In: *Nature Genetics* 44, p. 89. DOI: 10.1038/ng.1006https://www.nature.com/articles/ng.1006#supplementary-information.

Mayer, A., H. M. Landry, and L. S. Churchman (2017). "Pause & go: from the discovery of RNA polymerase pausing to its functional implications". In: *Curr Opin Cell Biol* 46, pp. 72–80. ISSN: 0955-0674. DOI: 10.1016/j.ceb.2017.03.002.

McLean, C. Y., D. Bristor, M. Hiller, S. L. Clarke, et al. (2010). "GREAT improves functional interpretation of cis-regulatory regions". In: *Nat Biotechnol* 28.5, pp. 495–501. ISSN: 1087-0156. DOI: 10.1038/nbt.1630.

McManus, S., A. Ebert, G. Salvagiotto, J. Medvedovic, et al. (2011). "The transcription factor Pax5 regulates its target genes by recruiting chromatin-modifying proteins in committed B cells". In: *Embo j* 30.12, pp. 2388–404. ISSN: 0261-4189. DOI: 10.1038/emboj.2011.140.

Mei, S., Q. Qin, Q. Wu, H. Sun, et al. (2017). "Cistrome Data Browser: a data portal for ChIP-Seq and chromatin accessibility data in human and mouse". In: *Nucleic Acids Res* 45.D1, pp. D658–d662. ISSN: 0305-1048. DOI: 10.1093/nar/gkw983.

Meireles-Filho, A. C. and A. Stark (2009). "Comparative genomics of gene regulation-conservation and divergence of cis-regulatory information". In: *Curr Opin Genet Dev* 19.6, pp. 565–70. ISSN: 0959-437x. DOI: 10.1016/j.gde.2009.10.006.

Melnikov, A., A. Murugan, X. Zhang, T. Tesileanu, et al. (2012). "Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay". In: *Nature Biotechnology* 30, p. 271. DOI: 10.1038/nbt.2137https://www.nature.com/articles/nbt.2137#supplementary-information.

Mercer, E. M., Y. C. Lin, C. Benner, S. Jhunjhunwala, et al. (2011). "Multilineage priming of enhancer repertoires precedes commitment to the B and myeloid cell lineages in hematopoietic progenitors". In: *Immunity* 35.3, pp. 413–25. ISSN: 1074-7613. DOI: 10.1016/j.immuni.2011.06.013.

Mifsud, B., F. Tavares-Cadete, A. N. Young, R. Sugar, et al. (2015). "Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C". In: *Nature Genetics* 47, p. 598. DOI: 10.1038/ng.3286https://www.nature.com/articles/ng.3286#supplementary-information.

Mohrs, M., C. M. Blankespoor, Z. E. Wang, G. G. Loots, et al. (2001). "Deletion of a coordinate regulator of type 2 cytokine expression in mice". In: *Nat Immunol* 2.9, pp. 842–7. ISSN: 1529-2908 (Print) 1529-2908. DOI: 10.1038/ni0901-842.

Moorthy, S. D., S. Davidson, V. M. Shchuka, G. Singh, et al. (2017). "Enhancers and super-enhancers have an equivalent regulatory role in embryonic stem cells through regulation of single or multiple genes". In: *Genome Res* 27.2, pp. 246–258. ISSN: 1088-9051. DOI: 10.1101/gr.210930.116.

Moorthy, S. D. and J. A. Mitchell (2016). "Generating CRISPR/Cas9 Mediated Monoallelic Deletions to Study Enhancer Function in Mouse Embryonic Stem Cells". In: *J Vis Exp* 110, e53552. ISSN: 1940-087x. DOI: 10.3791/53552.

Moreno-Indias, I., F. Cardona, F. J. Tinahones, and M. I. Queipo-Ortuno (2014). "Impact of the gut microbiota on the development of obesity and type 2 diabetes mellitus".

In: *Front Microbiol* 5, p. 190. ISSN: 1664-302X (Print) 1664-302x. DOI: 10.3389/fmicb.2014.00190.

Morgan, B., L. Sun, N. Avitahl, K. Andrikopoulos, et al. (1997). "Aiolos, a lymphoid restricted transcription factor that interacts with Ikaros to regulate lymphocyte differentiation". In: *Embo j* 16.8, pp. 2004–13. ISSN: 0261-4189 (Print) 0261-4189. DOI: 10.1093/emboj/16.8.2004.

Mori, Y., H. Kataoka, Y. Miura, M. Kawaguchi, et al. (2007). "Subcellular localization of ATBF1 regulates MUC5AC transcription in gastric cancer". In: *Int J Cancer* 121.2, pp. 241–7. ISSN: 0020-7136 (Print) 0020-7136. DOI: 10.1002/ijc.22654.

Morin, R. D., M. Mendez-Lago, A. J. Mungall, R. Goya, et al. (2011). "Frequent mutation of histone-modifying genes in non-Hodgkin lymphoma". In: *Nature* 476.7360, pp. 298–303. ISSN: 0028-0836. DOI: 10.1038/nature10351.

Morris, S. A., S. Baek, M. H. Sung, S. John, et al. (2014). "Overlapping chromatin-remodeling systems collaborate genome wide at dynamic chromatin transitions". In: *Nat Struct Mol Biol* 21.1, pp. 73–81. ISSN: 1545-9985. DOI: 10.1038/nsmb.2718.

Nacht, A. S., R. Ferrari, R. Zaurin, V. Scabia, et al. (2019). "C/EBP alpha mediates the growth inhibitory effect of progestins on breast cancer cells". In: *The EMBO Journal*, e101426. ISSN: 0261-4189. DOI: 10.15252/embj.2018101426.

Najafabadi, H. S., S. Mnaimneh, F. W. Schmitges, M. Garton, et al. (2015). "C2H2 zinc finger proteins greatly expand the human regulatory lexicon". In: *Nat Biotechnol* 33.5, pp. 555–62. ISSN: 1087-0156. DOI: 10.1038/nbt.3128.

Nakae, J., 3. Biggs W. H., T. Kitamura, W. K. Cavenee, et al. (2002). "Regulation of insulin action and pancreatic beta-cell function by mutated alleles of the gene encoding forkhead transcription factor Foxo1". In: *Nat Genet* 32.2, pp. 245–53. ISSN: 1061-4036 (Print) 1061-4036. DOI: 10.1038/ng890.

Nakae, J., M. Oki, and Y. Cao (2008). "The FoxO transcription factors and metabolic regulation". In: *FEBS Lett* 582.1, pp. 54–67. ISSN: 0014-5793 (Print) 0014-5793. DOI: 10.1016/j.febslet.2007.11.025.

NCD-Risk-Factor-Collaboration (2016). "Worldwide trends in diabetes since 1980: a pooled analysis of 751 population-based studies with 4.4 million participants". In: *Lancet* 387.10027, pp. 1513–30. ISSN: 0140-6736. DOI: 10.1016/s0140-6736(16)00618-8.

Nedergaard, J., N. Petrovic, E. M. Lindgren, A. Jacobsson, et al. (2005). "PPARgamma in the control of brown adipocyte differentiation". In: *Biochim Biophys Acta* 1740.2, pp. 293–304. ISSN: 0006-3002 (Print) 0006-3002. DOI: 10.1016/j.bbadis.2005.02.003.

Neph, S., J. Vierstra, A. B. Stergachis, A. P. Reynolds, et al. (2012). "An expansive human regulatory lexicon encoded in transcription factor footprints". In: *Nature* 489.7414, pp. 83–90. ISSN: 0028-0836. DOI: 10.1038/nature11212.

Nguyen, D. and T. Xu (2008). "The expanding role of mouse genetics for understanding human biology and disease". In: *Dis Model Mech* 1.1, pp. 56–66. ISSN: 1754-8403 (Print) 1754-8403. DOI: 10.1242/dmm.000232.

Le-Niculescu, H., S. D. Patel, M. Bhat, R. Kuczenski, et al. (2009). "Convergent functional genomics of genome-wide association data for bipolar disorder: comprehensive identification of candidate genes, pathways and mechanisms". In: *Am*

J Med Genet B Neuropsychiatr Genet 150b.2, pp. 155–81. ISSN: 1552-4841. DOI: 10.1002/ajmg.b.30887.

Niederriter, A. R., A. Varshney, S. C. Parker, and D. M. Martin (2015). "Super Enhancers in Cancers, Complex Disease, and Developmental Disorders". In: *Genes (Basel)* 6.4, pp. 1183–200. ISSN: 2073-4425 (Print) 2073-4425. DOI: 10.3390/genes6041183.

Nolan, P. M., J. Peters, M. Strivens, D. Rogers, et al. (2000). "A systematic, genome-wide, phenotype-driven mutagenesis programme for gene function studies in the mouse". In: *Nat Genet* 25.4, pp. 440–3. ISSN: 1061-4036 (Print) 1061-4036. DOI: 10.1038/78140.

Nord, A. S., M. J. Blow, C. Attanasio, J. A. Akiyama, et al. (2013). "Rapid and pervasive changes in genome-wide enhancer usage during mammalian development". In: *Cell* 155.7, pp. 1521–31. ISSN: 0092-8674. DOI: 10.1016/j.cell.2013.11.033.

Odom, D. T., R. D. Dowell, E. S. Jacobsen, W. Gordon, et al. (2007). "Tissue-specific transcriptional regulation has diverged significantly between human and mouse". In: *Nat Genet* 39.6, pp. 730–2. ISSN: 1061-4036 (Print) 1061-4036. DOI: 10.1038/ng2047.

Ogbourne, S. and T. M. Antalis (1998). "Transcriptional control and the role of silencers in transcriptional regulation in eukaryotes". In: *Biochem J* 331 (Pt 1), pp. 1–14. ISSN: 0264-6021 (Print) 0264-6021.

Ohba, S., X. He, H. Hojo, and A. P. McMahon (2015). "Distinct Transcriptional Programs Underlie Sox9 Regulation of the Mammalian Chondrocyte". In: *Cell Rep* 12.2, pp. 229–43. DOI: 10.1016/j.celrep.2015.06.013.

Ohno, S. (1972). "So much "junk" DNA in our genome". In: *Brookhaven Symp Biol* 23, pp. 366–70. ISSN: 0068-2799 (Print) 0068-2799.

Oldridge, D. A., A. C. Wood, N. Weichert-Leahey, I. Crimmins, et al. (2015). "Genetic predisposition to neuroblastoma mediated by a LMO1 super-enhancer polymorphism". In: *Nature* 528.7582, pp. 418–21. ISSN: 0028-0836. DOI: 10.1038/nature15540.

Ong, C. T. and V. G. Corces (2011). "Enhancer function: new insights into the regulation of tissue-specific gene expression". In: *Nat Rev Genet* 12.4, pp. 283–93. ISSN: 1471-0056. DOI: 10.1038/nrg2957.

Osterwalder, M., I. Barozzi, V. Tissières, Y. Fukuda-Yuzawa, et al. (2018). "Enhancer redundancy provides phenotypic robustness in mammalian development". In: *Nature* 554, p. 239. DOI: 10.1038/nature25461https://www.nature.com/articles/nature25461#supplementary-information.

Panagia, M., Y. C. Chen, H. H. Chen, L. Ernande, et al. (2016). "Functional and anatomical characterization of brown adipose tissue in heart failure with blood oxygen level dependent magnetic resonance". In: *NMR Biomed* 29.7, pp. 978–84. ISSN: 0952-3480. DOI: 10.1002/nbm.3557.

Parker, S. C., M. L. Stitzel, D. L. Taylor, J. M. Orozco, et al. (2013). "Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants". In: *Proc Natl Acad Sci U S A* 110.44, pp. 17921–6. ISSN: 0027-8424. DOI: 10.1073/pnas.1317023110.

Parsons, M. J., M. Brancaccio, S. Sethi, E. S. Maywood, et al. (2015). "The Regulatory Factor ZFHX3 Modifies Circadian Function in SCN via an AT Motif-Driven Axis". In: *Cell* 162.3, pp. 607–621. ISSN: 0092-8674. DOI: 10.1016/j.cell.2015.06.060.

Partch, C. L., C. B. Green, and J. S. Takahashi (2014). "Molecular architecture of the mammalian circadian clock". In: *Trends Cell Biol* 24.2, pp. 90–9. ISSN: 0962-8924. DOI: 10.1016/j.tcb.2013.07.002.

Pasquali, L., K. J. Gaulton, S. A. Rodriguez-Segui, L. Mularoni, et al. (2014). "Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants". In: *Nat Genet* 46.2, pp. 136–143. ISSN: 1061-4036. DOI: 10.1038/ng.2870.

Pasqualucci, L., D. Dominguez-Sola, A. Chiarenza, G. Fabbri, et al. (2011). "Inactivating mutations of acetyltransferase genes in B-cell lymphoma". In: *Nature* 471.7337, pp. 189–95. ISSN: 0028-0836. DOI: 10.1038/nature09730.

Patwardhan, R. P., J. B. Hiatt, D. M. Witten, M. J. Kim, et al. (2012). "Massively parallel functional dissection of mammalian enhancers in vivo". In: *Nat Biotechnol* 30.3, pp. 265–70. ISSN: 1087-0156. DOI: 10.1038/nbt.2136.

Peeters, J. G., S. J. Vervoort, S. C. Tan, G. Mijnheer, et al. (2015). "Inhibition of Super-Enhancer Activity in Autoinflammatory Site-Derived T Cells Reduces Disease-Associated Gene Expression". In: *Cell Rep* 12.12, pp. 1986–96. DOI: 10.1016/j.celrep.2015.08.046.

Pelish, H. E., B. B. Liau, I. I. Nitulescu, A. Tangpeerachaikul, et al. (2015). "Mediator kinase inhibition further activates super-enhancer-associated genes in AML". In: *Nature* 526, p. 273. DOI: 10.1038/nature14904https://www.nature.com/articles/nature14904#supplementary-information.

Pena-Castillo, L., M. Tasan, C. L. Myers, H. Lee, et al. (2008). "A critical assessment of Mus musculus gene function prediction using integrated genomic evidence". In: *Genome Biol* 9 Suppl 1, S2. ISSN: 1474-7596. DOI: 10.1186/gb-2008-9-s1-s2.

Pennacchio, L. A., N. Ahituv, A. M. Moses, S. Prabhakar, et al. (2006). "In vivo enhancer analysis of human conserved non-coding sequences". In: *Nature* 444.7118, pp. 499–502. ISSN: 0028-0836. DOI: 10.1038/nature05295.

Pennacchio, L. A., W. Bickmore, A. Dean, M. A. Nobrega, et al. (2013). "Enhancers: five essential questions". In: *Nat Rev Genet* 14.4, pp. 288–95. ISSN: 1471-0056. DOI: 10.1038/nrg3458.

Pfeifer, D., R. Kist, K. Dewar, K. Devon, et al. (1999). "Campomelic dysplasia translocation breakpoints are scattered over 1 Mb proximal to SOX9: evidence for an extended control region". In: *Am J Hum Genet* 65.1, pp. 111–24. ISSN: 0002-9297 (Print) 0002-9297. DOI: 10.1086/302455.

Pikkarainen, S., H. Tokola, R. Kerkela, and H. Ruskoaho (2004). "GATA transcription factors in the developing and adult heart". In: *Cardiovasc Res* 63.2, pp. 196–207. ISSN: 0008-6363 (Print) 0008-6363. DOI: 10.1016/j.cardiores.2004.03.025.

Pinero, J., A. Bravo, N. Queralt-Rosinach, A. Gutierrez-Sacristan, et al. (2017). "DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants". In: *Nucleic Acids Res* 45.D1, pp. D833–d839. ISSN: 0305-1048. DOI: 10.1093/nar/gkw943.

Pique-Regi, R., J. F. Degner, A. A. Pai, D. J. Gaffney, et al. (2011). "Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data". In: *Genome Res* 21.3, pp. 447–55. ISSN: 1088-9051. DOI: 10.1101/gr.112623.110.

Pomerantz, M. M., N. Ahmadiyeh, L. Jia, P. Herman, et al. (2009). "The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer". In: *Nat Genet* 41.8, pp. 882–4. ISSN: 1061-4036. DOI: 10.1038/ng.403.

Pott, S. and J. D. Lieb (2015). "What are super-enhancers?" In: *Nat Genet* 47.1, pp. 8–12. ISSN: 1061-4036. DOI: 10.1038/ng.3167.

Potter, P. K., M. R. Bowl, P. Jeyarajan, L. Wisby, et al. (2016). "Novel gene function revealed by mouse mutagenesis screens for models of age-related disease". In: *Nat Commun* 7, p. 12444. ISSN: 2041-1723. DOI: 10.1038/ncomms12444.

Poulsen, P., K. O. Kyvik, A. Vaag, and H. Beck-Nielsen (1999). "Heritability of type II (non-insulin-dependent) diabetes mellitus and abnormal glucose tolerance–a population-based twin study". In: *Diabetologia* 42.2, pp. 139–45. ISSN: 0012-186X (Print) 0012-186x.

Pradeepa, M. M., G. R. Grimes, Y. Kumar, G. Olley, et al. (2016). "Histone H3 globular domain acetylation identifies a new class of enhancers". In: *Nat Genet* 48.6, pp. 681–6. ISSN: 1061-4036 (Print). DOI: 10.1038/ng.3550.

Prosser, H. M., A. Bradley, J. E. Chesham, F. J. Ebling, et al. (2007). "Prokineticin receptor 2 (Prokr2) is essential for the regulation of circadian behavior by the suprachiasmatic nuclei". In: *Proc Natl Acad Sci U S A* 104.2, pp. 648–53. ISSN: 0027-8424 (Print) 0027-8424. DOI: 10.1073/pnas.0606884104.

Qi, Y., J. A. Ranish, X. Zhu, A. Krones, et al. (2008). "Atbf1 is required for the Pit1 gene early activation". In: *Proc Natl Acad Sci U S A* 105.7, pp. 2481–6. ISSN: 0027-8424. DOI: 10.1073/pnas.0712196105.

Qiu, W. Q. (2017). "Amylin and its G-protein-coupled receptor: A probable pathological process and drug target for Alzheimer's disease". In: *Neuroscience* 356, pp. 44–51. ISSN: 0306-4522. DOI: https://doi.org/10.1016/j.neuroscience.2017.05.024.

Quinlan, A. R. (2014). "BEDTools: The Swiss-Army Tool for Genome Feature Analysis". In: *Curr Protoc Bioinformatics* 47, pp. 11.12.1–34. ISSN: 1934-3396. DOI: 10.1002/0471250953.bi1112s47.

Rahimov, F., M. L. Marazita, A. Visel, M. E. Cooper, et al. (2008). "Disruption of an AP-2alpha binding site in an IRF6 enhancer is associated with cleft lip". In: *Nat Genet* 40.11, pp. 1341–7. ISSN: 1061-4036. DOI: 10.1038/ng.242.

Rajagopal, N., S. Srinivasan, K. Kooshesh, Y. Guo, et al. (2016). "High-throughput mapping of regulatory DNA". In: *Nat Biotechnol* 34.2, pp. 167–74. ISSN: 1087-0156. DOI: 10.1038/nbt.3468.

Rajakumari, S., J. Wu, J. Ishibashi, H. W. Lim, et al. (2013). "EBF2 determines and maintains brown adipocyte identity". In: *Cell Metab* 17.4, pp. 562–74. ISSN: 1550-4131. DOI: 10.1016/j.cmet.2013.01.015.

Ralph, M. R., R. G. Foster, F. C. Davis, and M. Menaker (1990). "Transplanted suprachiasmatic nucleus determines circadian period". In: *Science* 247.4945, pp. 975–8. ISSN: 0036-8075 (Print) 0036-8075.

Ramakrishnan, V. (2002). "Ribosome structure and the mechanism of translation". In: *Cell* 108.4, pp. 557–72. ISSN: 0092-8674 (Print) 0092-8674.

Rao, S. S., M. H. Huntley, N. C. Durand, E. K. Stamenova, et al. (2014). "A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping". In: *Cell* 159.7, pp. 1665–80. ISSN: 0092-8674. DOI: 10.1016/j.cell.2014.11.021.

Raymond, C. S., M. W. Murphy, M. G. O'Sullivan, V. J. Bardwell, et al. (2000). "Dmrt1, a gene related to worm and fly sexual regulators, is required for mammalian testis

differentiation". In: *Genes Dev* 14.20, pp. 2587–95. ISSN: 0890-9369 (Print) 0890-9369.

Reghunandanan, V. and R. Reghunandanan (2006). "Neurotransmitters of the suprachiasmatic nuclei". In: *J Circadian Rhythms* 4, p. 2. ISSN: 1740-3391. DOI: 10.1186/1740-3391-4-2.

Reilly, S. K., J. Yin, A. E. Ayoub, D. Emera, et al. (2015). "Evolutionary genomics. Evolutionary changes in promoter and enhancer activity during human corticogenesis". In: *Science* 347.6226, pp. 1155–9. ISSN: 0036-8075 (Print) 0036-8075. DOI: 10.1126/science.1260943.

Reimand, J., M. Kull, H. Peterson, J. Hansen, et al. (2007). "g:Profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments". In: *Nucleic Acids Res* 35.Web Server issue, W193–200. ISSN: 0305-1048 (Print). DOI: 10.1093/nar/gkm226.

Rhee, H. S. and B. F. Pugh (2011). "Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution". In: *Cell* 147.6, pp. 1408–19. ISSN: 0092-8674. DOI: 10.1016/j.cell.2011.11.013.

Roadmap Epigenomics Consortium, T., A. Kundaje, W. Meuleman, J. Ernst, et al. (2015). "Integrative analysis of 111 reference human epigenomes". In: *Nature* 518, p. 317. DOI: 10.1038/nature14248https://www.nature.com/articles/nature14248#supplementary-information.

Robinson, M. D., D. J. McCarthy, and G. K. Smyth (2010). "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data". In: *Bioinformatics* 26.1, pp. 139–40. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btp616.

Roelfsema, J. H., S. J. White, Y. Ariyurek, D. Bartholdi, et al. (2005). "Genetic heterogeneity in Rubinstein-Taybi syndrome: mutations in both the CBP and EP300 genes cause disease". In: *Am J Hum Genet* 76.4, pp. 572–80. ISSN: 0002-9297 (Print) 0002-9297. DOI: 10.1086/429130.

Ross-Innes, C. S., R. Stark, A. E. Teschendorff, K. A. Holmes, et al. (2012). "Differential oestrogen receptor binding is associated with clinical outcome in breast cancer". In: *Nature* 481.7381, pp. 389–393. ISSN: 0028-0836 1476-4687. DOI: 10.1038/nature10730.

Roth, F. P., J. D. Hughes, P. W. Estep, and G. M. Church (1998). "Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation". In: *Nat Biotechnol* 16.10, pp. 939–45. ISSN: 1087-0156 (Print) 1087-0156. DOI: 10.1038/nbt1098-939.

Safari-Alighiarloo, N., M. Taghizadeh, M. Rezaei-Tavirani, B. Goliaei, et al. (2014). "Protein-protein interaction networks (PPI) and complex diseases". In: *Gastroenterol Hepatol Bed Bench* 7.1, pp. 17–31. ISSN: 2008-2258 (Print) 2008-2258.

Sanchez-Castro, M., C. T. Gordon, F. Petit, A. S. Nord, et al. (2013). "Congenital heart defects in patients with deletions upstream of SOX9". In: *Hum Mutat* 34.12, pp. 1628–31. ISSN: 1059-7794. DOI: 10.1002/humu.22449.

Sandelin, A., P. Bailey, S. Bruce, P. G. Engström, et al. (2004). "Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes". In: *BMC Genomics* 5.1, p. 99. ISSN: 1471-2164. DOI: 10.1186/1471-2164-5-99.

Sanyal, A., B. R. Lajoie, G. Jain, and J. Dekker (2012). "The long-range interaction landscape of gene promoters". In: *Nature* 489.7414, pp. 109–13. ISSN: 0028-0836. DOI: 10.1038/nature11279.

Sarmento, O. F., P. A. Svingen, Y. Xiong, R. J. Xavier, et al. (2015). "A novel role for KLF14 in T regulatory cell differentiation". In: *Cell Mol Gastroenterol Hepatol* 1.2, 188–202.e4. ISSN: 2352-345X (Print) 2352-345x. DOI: 10.1016/j.jcmgh.2014.12.007.

Schirm, S., F. Weber, W. Schaffner, and B. Fleckenstein (1985). "A transcription enhancer in the Herpesvirus saimiri genome". In: *Embo j* 4.10, pp. 2669–74. ISSN: 0261-4189 (Print) 0261-4189.

Schofield, P. N., R. Hoehndorf, and G. V. Gkoutos (2012). "Mouse genetic and phenotypic resources for human genetics". In: *Hum Mutat* 33.5, pp. 826–36. ISSN: 1059-7794. DOI: 10.1002/humu.22077.

Schwessinger, R., M. C. Suciu, S. J. McGowan, J. Telenius, et al. (2017). "Sasquatch: predicting the impact of regulatory SNPs on transcription factor binding from cell- and tissue-specific DNase footprints". In: *Genome Res* 27.10, pp. 1730–1742. ISSN: 1088-9051 (Print) 1088-9051. DOI: 10.1101/gr.220202.117.

Scott, E. W., M. C. Simon, J. Anastasi, and H. Singh (1994a). "Requirement of transcription factor PU.1 in the development of multiple hematopoietic lineages". In: *Science* 265.5178, pp. 1573–7. ISSN: 0036-8075 (Print) 0036-8075.

— (1994b). "Requirement of transcription factor PU.1 in the development of multiple hematopoietic lineages". In: *Science* 265.5178, p. 1573.

Sengupta, S. and R. E. George (2017). "Super-Enhancer-Driven Transcriptional Dependencies in Cancer". In: *Trends Cancer* 3.4, pp. 269–281. ISSN: 2405-8025. DOI: 10.1016/j.trecan.2017.03.006.

Serna, I. L. de la, Y. Ohkawa, C. A. Berkes, D. A. Bergstrom, et al. (2005). "MyoD targets chromatin remodeling complexes to the myogenin locus prior to forming a stable DNA-bound complex". In: *Mol Cell Biol* 25.10, pp. 3997–4009. ISSN: 0270-7306 (Print) 0270-7306. DOI: 10.1128/mcb.25.10.3997-4009.2005.

Sertil, O., R. Kapoor, B. D. Cohen, N. Abramova, et al. (2003). "Synergistic repression of anaerobic genes by Mot3 and Rox1 in Saccharomyces cerevisiae". In: *Nucleic Acids Res* 31.20, pp. 5831–7. ISSN: 0305-1048.

Seruggia, D., A. Fernandez, M. Cantero, P. Pelczar, et al. (2015). "Functional validation of mouse tyrosinase non-coding regulatory DNA elements by CRISPR-Cas9-mediated mutagenesis". In: *Nucleic Acids Res* 43.10, pp. 4855–67. ISSN: 0305-1048. DOI: 10.1093/nar/gkv375.

Seyednasrollah, F., A. Laiho, and L. L. Elo (2015). "Comparison of software packages for detecting differential expression in RNA-seq studies". In: *Brief Bioinform* 16.1, pp. 59–70. ISSN: 1467-5463. DOI: 10.1093/bib/bbt086.

Shalem, O., N. E. Sanjana, E. Hartenian, X. Shi, et al. (2014). "Genome-scale CRISPR-Cas9 knockout screening in human cells". In: *Science* 343.6166, pp. 84–87. ISSN: 0036-8075. DOI: 10.1126/science.1247005.

Shannon, P., A. Markiel, O. Ozier, N. S. Baliga, et al. (2003). "Cytoscape: a software environment for integrated models of biomolecular interaction networks". In: *Genome Res* 13.11, pp. 2498–504. ISSN: 1088-9051 (Print) 1088-9051. DOI: 10.1101/gr.1239303.

Sharan, R., I. Ulitsky, and R. Shamir (2007). "Network-based prediction of protein function". In: *Mol Syst Biol* 3, p. 88. ISSN: 1744-4292. DOI: 10.1038/msb4100129.

Shen, L., N. Shao, X. Liu, and E. Nestler (2014). "ngs.plot: Quick mining and visualization of next-generation sequencing data by integrating genomic databases". In: *BMC Genomics* 15, p. 284. ISSN: 1471-2164. DOI: 10.1186/1471-2164-15-284.

Shen, Y., F. Yue, D. F. McCleary, Z. Ye, et al. (2012). "A map of the cis-regulatory sequences in the mouse genome". In: *Nature* 488.7409, pp. 116–20. ISSN: 0028-0836. DOI: 10.1038/nature11243.

Shi, J., W. A. Whyte, C. J. Zepeda-Mendoza, J. P. Milazzo, et al. (2013). "Role of SWI/SNF in acute leukemia maintenance and enhancer-mediated Myc regulation". In: *Genes Dev* 27.24, pp. 2648–62. ISSN: 0890-9369. DOI: 10.1101/gad.232710.113.

Shibata, Y., N. C. Sheffield, O. Fedrigo, C. C. Babbitt, et al. (2012). "Extensive Evolutionary Changes in Regulatory Element Activity during Human Origins Are Associated with Altered Gene Expression and Positive Selection". In: *PLOS Genetics* 8.6, e1002789. DOI: 10.1371/journal.pgen.1002789.

Shin, H. Y., M. Willi, K. HyunYoo, X. Zeng, et al. (2016). "Hierarchy within the mammary STAT5-driven Wap super-enhancer". In: *Nat Genet* 48.8, pp. 904–911. ISSN: 1061-4036. DOI: 10.1038/ng.3606.

Shlyueva, D., G. Stampfel, and A. Stark (2014). "Transcriptional enhancers: from properties to genome-wide predictions". In: *Nat Rev Genet* 15.4, pp. 272–86. ISSN: 1471-0056. DOI: 10.1038/nrg3682.

Shu, W., H. Chen, X. Bo, and S. Wang (2011). "Genome-wide analysis of the relationships between DNaseI HS, histone modifications and gene expression reveals distinct modes of chromatin domains". In: *Nucleic Acids Res* 39.17, pp. 7428–43. ISSN: 0305-1048 (Print). DOI: 10.1093/nar/gkr443.

Siepel, A., G. Bejerano, J. S. Pedersen, A. S. Hinrichs, et al. (2005). "Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes". In: *Genome Res* 15.8, pp. 1034–50. ISSN: 1088-9051 (Print) 1088-9051. DOI: 10.1101/gr.3715005.

Siersbæk, R., A. Rabiee, R. Nielsen, S. Sidoli, et al. (2014). "Transcription Factor Cooperativity in Early Adipogenic Hotspots and Super-Enhancers". In: *Cell Reports* 7.5, pp. 1443–1455. ISSN: 2211-1247. DOI: 10.1016/j.celrep.2014.04.042.

Slatkin, M. (2008). "Linkage disequilibrium–understanding the evolutionary past and mapping the medical future". In: *Nat Rev Genet* 9.6, pp. 477–85. ISSN: 1471-0056. DOI: 10.1038/nrg2361.

Small, K. S., A. K. Hedman, E. Grundberg, A. C. Nica, et al. (2011). "Identification of an imprinted master trans regulator at the KLF14 locus related to multiple metabolic phenotypes". In: *Nat Genet* 43.6, pp. 561–4. ISSN: 1061-4036. DOI: 10.1038/ng.833.

Small, K. S., M. Todorčević, M. Civelek, J. S. El-Sayed Moustafa, et al. (2018). "Regulatory variants at KLF14 influence type 2 diabetes risk via a female-specific effect on adipocyte size and body composition". In: *Nature Genetics* 50.4, pp. 572–580. ISSN: 1546-1718. DOI: 10.1038/s41588-018-0088-x.

Smemo, S., J. J. Tena, K. H. Kim, E. R. Gamazon, et al. (2014). "Obesity-associated variants within FTO form long-range functional connections with IRX3". In: *Nature* 507.7492, pp. 371–5. ISSN: 0028-0836. DOI: 10.1038/nature13138.

Smith, C. L., C. A. Goldsmith, and J. T. Eppig (2005). "The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information". In: *Genome Biol* 6.1, R7. ISSN: 1474-7596. DOI: 10.1186/gb-2004-6-1-r7.

Soldner, F., Y. Stelzer, C. S. Shivalila, B. J. Abraham, et al. (2016). "Parkinson-associated risk variant in distal enhancer of alpha-synuclein modulates target gene expression". In: *Nature* 533.7601, pp. 95–9. ISSN: 0028-0836. DOI: 10.1038/nature17939.

Song, L., Z. Zhang, L. L. Grasfeder, A. P. Boyle, et al. (2011). "Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity". In: *Genome Res* 21.10, pp. 1757–67. ISSN: 1088-9051. DOI: 10.1101/gr.121541.111.

Sotelo, J., D. Esposito, M. A. Duhagon, K. Banfield, et al. (2010). "Long-range enhancers on 8q24 regulate c-Myc". In: *Proc Natl Acad Sci U S A* 107.7, pp. 3001–5. ISSN: 0027-8424. DOI: 10.1073/pnas.0906067107.

Spandidos, D. A. and N. M. Wilkie (1983). "Host-specificities of papillomavirus, Moloney murine sarcoma virus and simian virus 40 enhancer sequences". In: *Embo j* 2.7, pp. 1193–9. ISSN: 0261-4189 (Print) 0261-4189.

Spirin, V. and L. A. Mirny (2003). "Protein complexes and functional modules in molecular networks". In: *Proc Natl Acad Sci U S A* 100.21, pp. 12123–8. ISSN: 0027-8424 (Print) 0027-8424. DOI: 10.1073/pnas.2032324100.

Spitz, F. and E. E. Furlong (2012). "Transcription factors: from enhancer binding to developmental control". In: *Nat Rev Genet* 13.9, pp. 613–26. ISSN: 1471-0056. DOI: 10.1038/nrg3207.

Stark, C., B. J. Breitkreutz, T. Reguly, L. Boucher, et al. (2006). "BioGRID: a general repository for interaction datasets". In: *Nucleic Acids Res* 34.Database issue, pp. D535–9. ISSN: 0305-1048. DOI: 10.1093/nar/gkj109.

Steensel, B. van and S. Henikoff (2000). "Identification of in vivo DNA targets of chromatin proteins using tethered dam methyltransferase". In: *Nat Biotechnol* 18.4, pp. 424–8. ISSN: 1087-0156 (Print) 1087-0156. DOI: 10.1038/74487.

Stephens, J. M., R. F. Morrison, and P. F. Pilch (1996). "The expression and regulation of STATs during 3T3-L1 adipocyte differentiation". In: *J Biol Chem* 271.18, pp. 10441–4. ISSN: 0021-9258 (Print) 0021-9258.

Stone, N. R., C. A. Gifford, R. Thomas, K. J. B. Pratt, et al. (2019). "Context-Specific Transcription Factor Functions Regulate Epigenomic and Transcriptional Dynamics during Cardiac Reprogramming". In: *Cell Stem Cell* 25.1, 87–102.e9. ISSN: 1934-5909. DOI: https://doi.org/10.1016/j.stem.2019.06.012.

Stormo, G. D. (2000). "DNA binding sites: representation and discovery". In: *Bioinformatics* 16.1, pp. 16–23. ISSN: 1367-4803 (Print) 1367-4803.

Sun, F. L. and S. C. Elgin (1999). "Putting boundaries on silence". In: *Cell* 99.5, pp. 459–62. ISSN: 0092-8674 (Print) 0092-8674.

Sun, J., G. Matthias, M. J. Mihatsch, K. Georgopoulos, et al. (2003). "Lack of the transcriptional coactivator OBF-1 prevents the development of systemic lupus erythematosus-like phenotypes in Aiolos mutant mice". In: *J Immunol* 170.4, pp. 1699–706. ISSN: 0022-1767 (Print) 0022-1767.

Sun, X., X. Fu, J. Li, C. Xing, et al. (2012). "Heterozygous deletion of Atbf1 by the Cre-loxP system in mice causes preweaning mortality". In: *Genesis* 50.11, pp. 819–27. ISSN: 1526-954x. DOI: 10.1002/dvg.22041.

Sur, I. K., O. Hallikas, A. Vaharautio, J. Yan, et al. (2012). "Mice lacking a Myc enhancer that includes human SNP rs6983267 are resistant to intestinal tumors". In: *Science* 338.6112, pp. 1360–3. ISSN: 0036-8075. DOI: 10.1126/science.1228606.

Suzuki, H. I., R. A. Young, and P. A. Sharp (2017). "Super-Enhancer-Mediated RNA Processing Revealed by Integrative MicroRNA Network Analysis". In: *Cell* 168.6, 1000–1014.e15. ISSN: 0092-8674. DOI: 10.1016/j.cell.2017.02.015.

Swift, J. and G. Coruzzi (2017). "A Matter of Time - How Transient Transcription Factor Interactions Create Dynamic Gene Regulatory Networks". In: *Biochim Biophys Acta* 1860.1, pp. 75–83. ISSN: 0006-3002 (Print). DOI: 10.1016/j.bbagrm.2016.08.007.

Taher, L., D. M. McGaughey, S. Maragh, I. Aneas, et al. (2011). "Genome-wide identification of conserved regulatory function in diverged sequences". In: *Genome Res* 21.7, pp. 1139–49. ISSN: 1088-9051. DOI: 10.1101/gr.119016.110.

Tasan, M., W. Tian, D. P. Hill, F. D. Gibbons, et al. (2008). "An en masse phenotype and function prediction system for Mus musculus". In: *Genome Biol* 9 Suppl 1, S8. ISSN: 1474-7596. DOI: 10.1186/gb-2008-9-s1-s8.

Tasdemir, N., A. Banito, J. S. Roe, D. Alonso-Curbelo, et al. (2016). "BRD4 Connects Enhancer Remodeling to Senescence Immune Surveillance". In: *Cancer Discov* 6.6, pp. 612–29. ISSN: 2159-8274. DOI: 10.1158/2159-8290.cd-16-0217.

Taylor, R. (2013). "Type 2 diabetes: etiology and reversibility". In: *Diabetes Care* 36.4, pp. 1047–55. ISSN: 0149-5992. DOI: 10.2337/dc12-1805.

Teppo, S., S. Laukkanen, T. Liuksiala, J. Nordlund, et al. (2016). "Genome-wide repression of eRNA and target gene loci by the ETV6-RUNX1 fusion in acute leukemia". In: *Genome Res* 26.11, pp. 1468–1477. ISSN: 1088-9051. DOI: 10.1101/gr.193649.115.

Teslovich, T. M., K. Musunuru, A. V. Smith, A. C. Edmondson, et al. (2010). "Biological, clinical and population relevance of 95 loci for blood lipids". In: *Nature* 466.7307, pp. 707–13. ISSN: 0028-0836. DOI: 10.1038/nature09270.

Thathiah, A. and B. De Strooper (2011). "The role of G protein-coupled receptors in the pathology of Alzheimer's disease". In: *Nat Rev Neurosci* 12.2, pp. 73–87. ISSN: 1471-003x. DOI: 10.1038/nrn2977.

Thaung, C., K. West, B. J. Clark, L. McKie, et al. (2002). "Novel ENU-induced eye mutations in the mouse: models for human eye disease". In: *Hum Mol Genet* 11.7, pp. 755–67. ISSN: 0964-6906 (Print) 0964-6906.

Thoonen, R., L. Ernande, J. Cheng, Y. Nagasaka, et al. (2015). "Functional brown adipose tissue limits cardiomyocyte injury and adverse remodeling in catecholamine-induced cardiomyopathy". In: *J Mol Cell Cardiol* 84, pp. 202–11. ISSN: 0022-2828. DOI: 10.1016/j.yjmcc.2015.05.002.

Thoonen, R., A. G. Hindle, and M. Scherrer-Crosbie (2016). "Brown adipose tissue: The heat is on the heart". In: *American Journal of Physiology-Heart and Circulatory Physiology* 310.11, H1592–H1605. DOI: 10.1152/ajpheart.00698.2015.

Thurman, R. E., E. Rynes, R. Humbert, J. Vierstra, et al. (2012). "The accessible chromatin landscape of the human genome". In: *Nature* 489, p. 75. DOI: 10.1038/

nature11232https : / / www . nature . com / articles / nature11232 # supplementary - information.

Tippens, N. D., A. Vihervaara, and J. T. Lis (2018). "Enhancer transcription: what, where, when, and why?" In: *Genes Dev* 32.1, pp. 1–3. ISSN: 0890-9369. DOI: 10.1101/gad.311605.118.

Tolhuis, B., R. J. Palstra, E. Splinter, F. Grosveld, et al. (2002). "Looping and interaction between hypersensitive sites in the active beta-globin locus". In: *Mol Cell* 10.6, pp. 1453–65. ISSN: 1097-2765 (Print) 1097-2765.

Touzet, H. and J.-S. Varré (2007). "Efficient and accurate P-value computation for Position Weight Matrices". In: *Algorithms for molecular biology : AMB* 2, pp. 15–15. ISSN: 1748-7188. DOI: 10.1186/1748-7188-2-15.

Trapnell, C., A. Roberts, L. Goff, G. Pertea, et al. (2012). "Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks". In: *Nat Protoc* 7.3, pp. 562–78. ISSN: 1750-2799. DOI: 10.1038/nprot.2012.016.

Trudel, M. and F. Costantini (1987). "A 3' enhancer contributes to the stage-specific expression of the human beta-globin gene". In: *Genes Dev* 1.9, pp. 954–61. ISSN: 0890-9369 (Print) 0890-9369.

Ukai, H. and H. R. Ueda (2010). "Systems biology of mammalian circadian clocks". In: *Annu Rev Physiol* 72, pp. 579–603. ISSN: 0066-4278. DOI: 10.1146/annurev-physiol-073109-130051.

Vahedi, G., Y. Kanno, Y. Furumoto, K. Jiang, et al. (2015). "Super-enhancers delineate disease-associated regulatory nodes in T cells". In: *Nature* 520.7548, pp. 558–62. ISSN: 0028-0836. DOI: 10.1038/nature14154.

Vakoc, C. R., D. L. Letting, N. Gheldof, T. Sawado, et al. (2005). "Proximity among distant regulatory elements at the beta-globin locus requires GATA-1 and FOG-1". In: *Mol Cell* 17.3, pp. 453–62. ISSN: 1097-2765 (Print) 1097-2765. DOI: 10.1016/j.molcel.2004.12.028.

Vernimmen, D. and W. A. Bickmore (2015). "The Hierarchy of Transcriptional Activation: From Enhancer to Promoter". In: *Trends Genet* 31.12, pp. 696–708. ISSN: 0168-9525 (Print) 0168-9525. DOI: 10.1016/j.tig.2015.10.004.

Villar, D., C. Berthelot, S. Aldridge, T. F. Rayner, et al. (2015). "Enhancer evolution across 20 mammalian species". In: *Cell* 160.3, pp. 554–66. ISSN: 0092-8674. DOI: 10.1016/j.cell.2015.01.006.

Villiers, J. de, L. Olson, C. Tyndall, and W. Schaffner (1982). "Transcriptional 'enhancers' from SV40 and polyoma virus show a cell type preference". In: *Nucleic Acids Res* 10.24, pp. 7965–76. ISSN: 0305-1048 (Print) 0305-1048.

Visel, A., M. J. Blow, Z. Li, T. Zhang, et al. (2009). "ChIP-seq accurately predicts tissue-specific activity of enhancers". In: *Nature* 457.7231, pp. 854–8. ISSN: 0028-0836. DOI: 10.1038/nature07730.

Visel, A., S. Minovitsky, I. Dubchak, and L. A. Pennacchio (2007). "VISTA Enhancer Browser–a database of tissue-specific human enhancers". In: *Nucleic Acids Res* 35.Database issue, pp. D88–92. ISSN: 0305-1048. DOI: 10.1093/nar/gkl822.

Visscher, P. M., N. R. Wray, Q. Zhang, P. Sklar, et al. (2017). "10 Years of GWAS Discovery: Biology, Function, and Translation". In: *Am J Hum Genet* 101.1, pp. 5–22. ISSN: 0002-9297. DOI: 10.1016/j.ajhg.2017.06.005.

Voight, B. F., L. J. Scott, V. Steinthorsdottir, A. P. Morris, et al. (2010). "Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis". In: *Nat Genet* 42.7, pp. 579–89. ISSN: 1061-4036. DOI: 10.1038/ng.609.

Vorontsov, I. E., A. D. Fedorova, I. S. Yevshin, R. N. Sharipov, et al. (2018). "Genome-wide map of human and mouse transcription factor binding sites aggregated from ChIP-Seq data". In: *BMC Res Notes* 11.1, p. 756. ISSN: 1756-0500. DOI: 10.1186/s13104-018-3856-x.

Wahl, M. C., C. L. Will, and R. Luhrmann (2009). "The spliceosome: design principles of a dynamic RNP machine". In: *Cell* 136.4, pp. 701–18. ISSN: 0092-8674. DOI: 10.1016/j.cell.2009.02.009.

Walker, B. A., C. P. Wardell, A. Brioli, E. Boyle, et al. (2014). "Translocations at 8q24 juxtapose MYC with genes that harbor superenhancers resulting in overexpression and poor prognosis in myeloma patients". In: *Blood Cancer J* 4, e191. ISSN: 2044-5385 (Print) 2044-5385. DOI: 10.1038/bcj.2014.13.

Wang, H., G. D. Ferguson, V. V. Pineda, P. E. Cundiff, et al. (2004). "Overexpression of type-1 adenylyl cyclase in mouse forebrain enhances recognition memory and LTP". In: *Nat Neurosci* 7.6, pp. 635–42. ISSN: 1097-6256 (Print) 1097-6256. DOI: 10.1038/nn1248.

Wang, H. and M. Zhang (2012). "The role of Ca(2)(+)-stimulated adenylyl cyclases in bidirectional synaptic plasticity and brain function". In: *Rev Neurosci* 23.1, pp. 67–78. ISSN: 0334-1763 (Print) 0334-1763. DOI: 10.1515/revneuro-2011-0063.

Wang, J. H., N. Avitahl, A. Cariappa, C. Friedrich, et al. (1998). "Aiolos regulates B cell activation and maturation to effector state". In: *Immunity* 9.4, pp. 543–53. ISSN: 1074-7613 (Print) 1074-7613.

Wang, Q., J. S. Carroll, and M. Brown (2005). "Spatial and temporal recruitment of androgen receptor and its coactivators involves chromosomal looping and polymerase tracking". In: *Mol Cell* 19.5, pp. 631–42. ISSN: 1097-2765 (Print) 1097-2765. DOI: 10.1016/j.molcel.2005.07.018.

Wang, Z., C. Zang, K. Cui, D. E. Schones, et al. (2009). "Genome-wide mapping of HATs and HDACs reveals distinct functions in active and inactive genes". In: *Cell* 138.5, pp. 1019–31. ISSN: 0092-8674. DOI: 10.1016/j.cell.2009.06.049.

Wasserman, W. W. and A. Sandelin (2004). "Applied bioinformatics for the identification of regulatory elements". In: *Nat Rev Genet* 5.4, pp. 276–87. ISSN: 1471-0056 (Print) 1471-0056. DOI: 10.1038/nrg1315.

Wei, Y., S. Zhang, S. Shang, B. Zhang, et al. (2016). "SEA: a super-enhancer archive". In: *Nucleic Acids Res* 44.D1, pp. D172–9. ISSN: 0305-1048. DOI: 10.1093/nar/gkv1243.

Weintraub, H. and M. Groudine (1976). "Chromosomal subunits in active genes have an altered conformation". In: *Science* 193.4256, pp. 848–56. ISSN: 0036-8075 (Print) 0036-8075.

Welter, D., J. MacArthur, J. Morales, T. Burdett, et al. (2014). "The NHGRI GWAS Catalog, a curated resource of SNP-trait associations". In: *Nucleic Acids Res* 42.Database issue, pp. D1001–6. ISSN: 0305-1048. DOI: 10.1093/nar/gkt1229.

West, A. G., M. Gaszner, and G. Felsenfeld (2002). "Insulators: many functions, many mechanisms". In: *Genes Dev* 16.3, pp. 271–88. ISSN: 0890-9369 (Print) 0890-9369. DOI: 10.1101/gad.954702.

Whitington, T., M. C. Frith, J. Johnson, and T. L. Bailey (2011). "Inferring transcription factor complexes from ChIP-seq data". In: *Nucleic Acids Res* 39.15, e98. ISSN: 0305-1048. DOI: 10.1093/nar/gkr341.

Whyte, W. A., D. A. Orlando, D. Hnisz, B. J. Abraham, et al. (2013). "Master Transcription Factors and Mediator Establish Super-Enhancers at Key Cell Identity Genes". In: *Cell* 153.2, pp. 307–319. ISSN: 0092-8674. DOI: 10.1016/j.cell.2013.03.035.

Willemsen, G., K. J. Ward, C. G. Bell, K. Christensen, et al. (2015). "The Concordance and Heritability of Type 2 Diabetes in 34,166 Twin Pairs From International Twin Registers: The Discordant Twin (DISCOTWIN) Consortium". In: *Twin Res Hum Genet* 18.6, pp. 762–71. ISSN: 1832-4274 (Print) 1832-4274. DOI: 10.1017/thg.2015.83.

Winandy, S., P. Wu, and K. Georgopoulos (1995). "A dominant mutation in the Ikaros gene leads to rapid development of leukemia and lymphoma". In: *Cell* 83.2, pp. 289–99. ISSN: 0092-8674 (Print) 0092-8674.

Wingender, E., P. Dietze, H. Karas, and R. Knüppel (1996). "TRANSFAC: a database on transcription factors and their DNA binding sites". In: *Nucleic Acids Res* 24.1, pp. 238–41. ISSN: 0305-1048 (Print) 0305-1048.

Wittkopp, P. J. and G. Kalay (2011). "Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence". In: *Nature Reviews Genetics* 13, p. 59. DOI: 10.1038/nrg3095https://www.nature.com/articles/nrg3095#supplementary-information.

Wong, R. W. J., P. C. T. Ngoc, W. Z. Leong, A. W. Y. Yam, et al. (2017). "Enhancer profiling identifies critical cancer genes and characterizes cell identity in adult T-cell leukemia". In: *Blood* 130.21, pp. 2326–2338. ISSN: 0006-4971. DOI: 10.1182/blood-2017-06-792184.

Worsley Hunt, R. and W. W. Wasserman (2014). "Non-targeted transcription factors motifs are a systemic component of ChIP-seq datasets". In: *Genome Biol* 15.7. ISSN: 1465-6906 (Print). DOI: 10.1186/s13059-014-0412-4.

Wyce, A., G. Ganji, K. N. Smitheman, C. W. Chung, et al. (2013). "BET inhibition silences expression of MYCN and BCL2 and induces cytotoxicity in neuroblastoma tumor models". In: *PLoS One* 8.8, e72967. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0072967.

Xi, H., H. P. Shulha, J. M. Lin, T. R. Vales, et al. (2007). "Identification and characterization of cell type-specific and ubiquitous chromatin regulatory structures in the human genome". In: *PLoS Genet* 3.8, e136. ISSN: 1553-7390. DOI: 10.1371/journal.pgen.0030136.

Xi, Y., W. Shen, L. Ma, M. Zhao, et al. (2016). "HMGA2 promotes adipogenesis by activating C/EBPbeta-mediated expression of PPARgamma". In: *Biochem Biophys Res Commun* 472.4, pp. 617–23. ISSN: 0006-291x. DOI: 10.1016/j.bbrc.2016.03.015.

Xia, Z., E. J. Choi, F. Wang, C. Blazynski, et al. (1993). "Type I calmodulin-sensitive adenylyl cyclase is neural specific". In: *J Neurochem* 60.1, pp. 305–11. ISSN: 0022-3042 (Print) 0022-3042.

Yan, J., M. Enge, T. Whitington, K. Dave, et al. (2013). "Transcription factor binding in human cells occurs in dense clusters formed around cohesin anchor sites". In: *Cell* 154.4, pp. 801–13. ISSN: 0092-8674. DOI: 10.1016/j.cell.2013.07.034.

Yanai, I., H. Benjamin, M. Shmoish, V. Chalifa-Caspi, et al. (2005). "Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification". In: *Bioinformatics* 21.5, pp. 650–9. ISSN: 1367-4803 (Print) 1367-4803. DOI: 10.1093/bioinformatics/bti042.

Yanez-Cuna, J. O., H. Q. Dinh, E. Z. Kvon, D. Shlyueva, et al. (2012). "Uncovering cis-regulatory sequence requirements for context-specific transcription factor binding". In: *Genome Res* 22.10, pp. 2018–30. ISSN: 1088-9051. DOI: 10.1101/gr.132811.111.

Yang, W. M., H. J. Jeong, S. Y. Park, and W. Lee (2014). "Induction of miR-29a by saturated fatty acids impairs insulin signaling and glucose uptake through translational repression of IRS-1 in myocytes". In: *FEBS Lett* 588.13, pp. 2170–6. ISSN: 0014-5793. DOI: 10.1016/j.febslet.2014.05.011.

Yang, X. F., P. Fang, S. Meng, M. Jan, et al. (2009). "THE FORKHEAD TRANSCRIPTION FACTORS ARE IMPORTANT IN REGULATING VASCULAR PATHOLOGY, DIABETES AND REGULATORY T CELLS". In: *Front Biosci (Schol Ed)* 1, pp. 420–36. ISSN: 1945-0516 (Print).

Yao, L., B. P. Berman, and P. J. Farnham (2015). "Demystifying the secret mission of enhancers: linking distal regulatory elements to target genes". In: *Crit Rev Biochem Mol Biol* 50.6, pp. 550–73. ISSN: 1040-9238. DOI: 10.3109/10409238.2015.1087961.

Yasuda, H., A. Mizuno, T. Tamaoki, and T. Morinaga (1994). "ATBF1, a multiple-homeodomain zinc finger protein, selectively down-regulates AT-rich elements of the human alpha-fetoprotein gene". In: *Mol Cell Biol* 14.2, pp. 1395–401. ISSN: 0270-7306 (Print) 0270-7306.

Yevshin, I., R. Sharipov, S. Kolmykov, Y. Kondrakhin, et al. (2018). "GTRD: a database on gene transcription regulation-2019 update". In: *Nucleic Acids Res*. ISSN: 0305-1048. DOI: 10.1093/nar/gky1128.

Yuan, G. C., Y. J. Liu, M. F. Dion, M. D. Slack, et al. (2005). "Genome-scale identification of nucleosome positions in S. cerevisiae". In: *Science* 309.5734, pp. 626–30. ISSN: 0036-8075. DOI: 10.1126/science.1112178.

Yuan, Y., Y. Xu, J. Xu, R. L. Ball, et al. (2012). "Predicting the lethal phenotype of the knockout mouse by integrating comprehensive genomic data". In: *Bioinformatics* 28.9, pp. 1246–52. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bts120.

Yue, F., Y. Cheng, A. Breschi, J. Vierstra, et al. (2014). "A comparative encyclopedia of DNA elements in the mouse genome". In: *Nature* 515.7527, pp. 355–64. ISSN: 0028-0836. DOI: 10.1038/nature13992.

Zabidi, M. A., C. D. Arnold, K. Schernhuber, M. Pagani, et al. (2015). "Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation". In: *Nature* 518.7540, pp. 556–9. ISSN: 0028-0836. DOI: 10.1038/nature13994.

Zambelli, F., G. Pesole, and G. Pavesi (2009). "Pscan: finding over-represented transcription factor binding site motifs in sequences from co-regulated or co-expressed genes". In: *Nucleic Acids Res* 37.Web Server issue, W247–52. ISSN: 0305-1048. DOI: 10.1093/nar/gkp464.

Zeitlinger, J., R. P. Zinzen, A. Stark, M. Kellis, et al. (2007). "Whole-genome ChIP-chip analysis of Dorsal, Twist, and Snail suggests integration of diverse patterning processes in the Drosophila embryo". In: *Genes Dev* 21.4, pp. 385–90. ISSN: 0890-9369 (Print) 0890-9369. DOI: 10.1101/gad.1509607.

Zeng, L. and M. M. Zhou (2002). "Bromodomain: an acetyl-lysine binding domain". In: *FEBS Lett* 513.1, pp. 124–8. ISSN: 0014-5793 (Print) 0014-5793.

Zentner, G. E., P. J. Tesar, and P. C. Scacheri (2011). "Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions". In: *Genome Res* 21.8, pp. 1273–83. ISSN: 1088-9051. DOI: 10.1101/gr.122382.111.

Zhang, T., N. Kwiatkowski, C. M. Olson, S. E. Dixon-Clarke, et al. (2016). "Covalent targeting of remote cysteine residues to develop CDK12 and CDK13 inhibitors". In: *Nat Chem Biol* 12.10, pp. 876–84. ISSN: 1552-4450. DOI: 10.1038/nchembio.2166.

Zhang, X., P. S. Choi, J. M. Francis, M. Imielinski, et al. (2016). "Identification of focally amplified lineage-specific super-enhancers in human epithelial cancers". In: *Nat Genet* 48.2, pp. 176–82. ISSN: 1061-4036. DOI: 10.1038/ng.3470.

Zhang, Y., T. Liu, C. A. Meyer, J. Eeckhoute, et al. (2008). "Model-based Analysis of ChIP-Seq (MACS)". In: *Genome Biology* 9.9, R137. ISSN: 1474-760X. DOI: 10.1186/gb-2008-9-9-r137.

Zhao, J., Y. Deng, Z. Jiang, and H. Qing (2016). "G Protein-Coupled Receptors (GPCRs) in Alzheimer's Disease: A Focus on BACE1 Related GPCRs". In: *Front Aging Neurosci* 8, p. 58. ISSN: 1663-4365 (Print) 1663-4365. DOI: 10.3389/fnagi.2016.00058.

Zhao, Z., G. Tavoosidana, M. Sjölinder, A. Göndör, et al. (2006). "Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions". In: *Nature Genetics* 38, p. 1341. DOI: 10.1038/ng1891https://www.nature.com/articles/ng1891#supplementary-information.

Zhou, H., S. C. Schmidt, S. Jiang, B. Willox, et al. (2015). "Epstein-Barr virus oncoprotein super-enhancers control B cell growth". In: *Cell Host Microbe* 17.2, pp. 205–16. ISSN: 1931-3128. DOI: 10.1016/j.chom.2014.12.013.

Zhu, C., L. Li, Z. Zhang, M. Bi, et al. (2019). "A Non-canonical Role of YAP/TEAD Is Required for Activation of Estrogen-Regulated Enhancers in Breast Cancer". In: *Molecular Cell*. ISSN: 1097-2765. DOI: https://doi.org/10.1016/j.molcel.2019.06.010.

Zlotorynski, E. (2018). "Developmental enhancers in action". In: *Nature Reviews Molecular Cell Biology* 19, p. 210. DOI: 10.1038/nrm.2018.15.

Zschaler, J., D. Schlorke, and J. Arnhold (2014). "Differences in innate immune response between man and mouse". In: *Crit Rev Immunol* 34.5, pp. 433–54. ISSN: 1040-8401 (Print) 1040-8401.

Appendix A

Supplementary figures

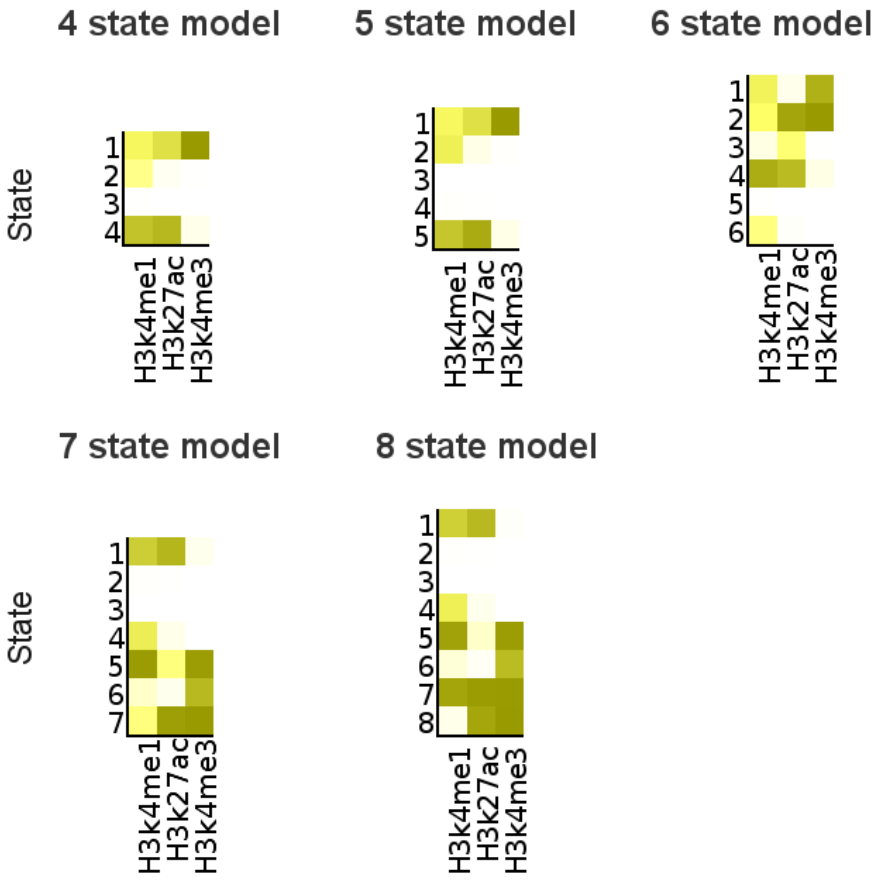


Fig. A.1 ChromHMM models with different number of chromatin states. Heatmaps display the ChromHMM models with different number of chromatin states used to segment the mouse genome.

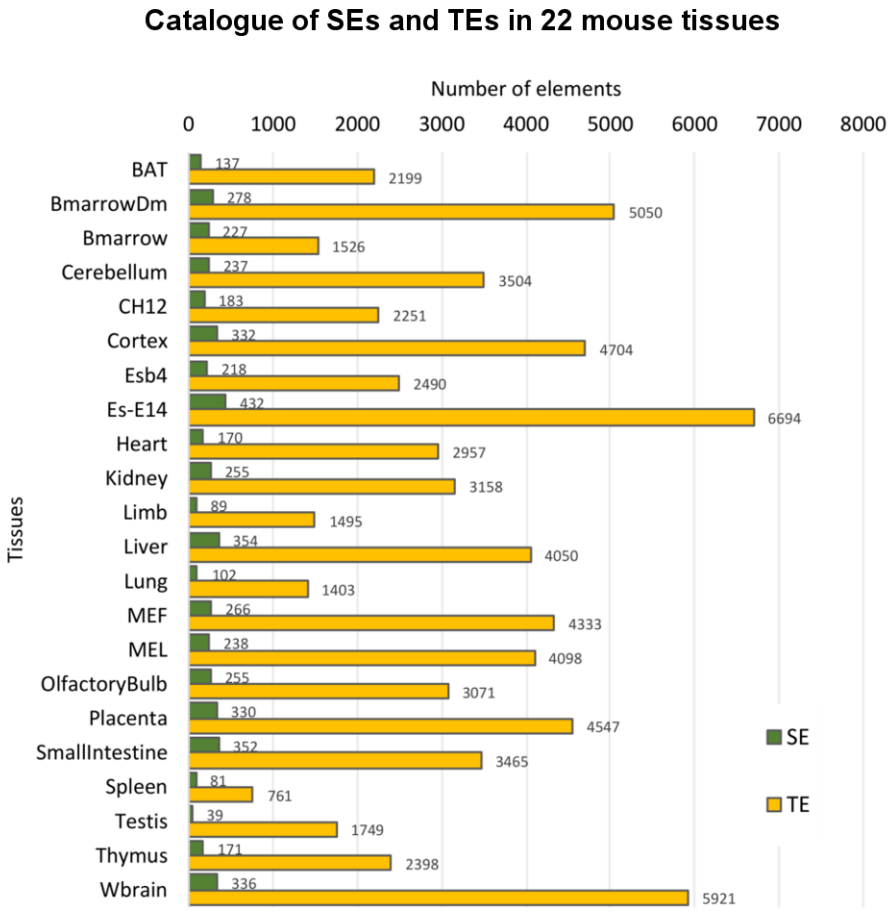


Fig. A.2 SEs and TEs identified in 22 mouse tissues. Bar plot displaying the number of SEs and TEs identified in each tissue.

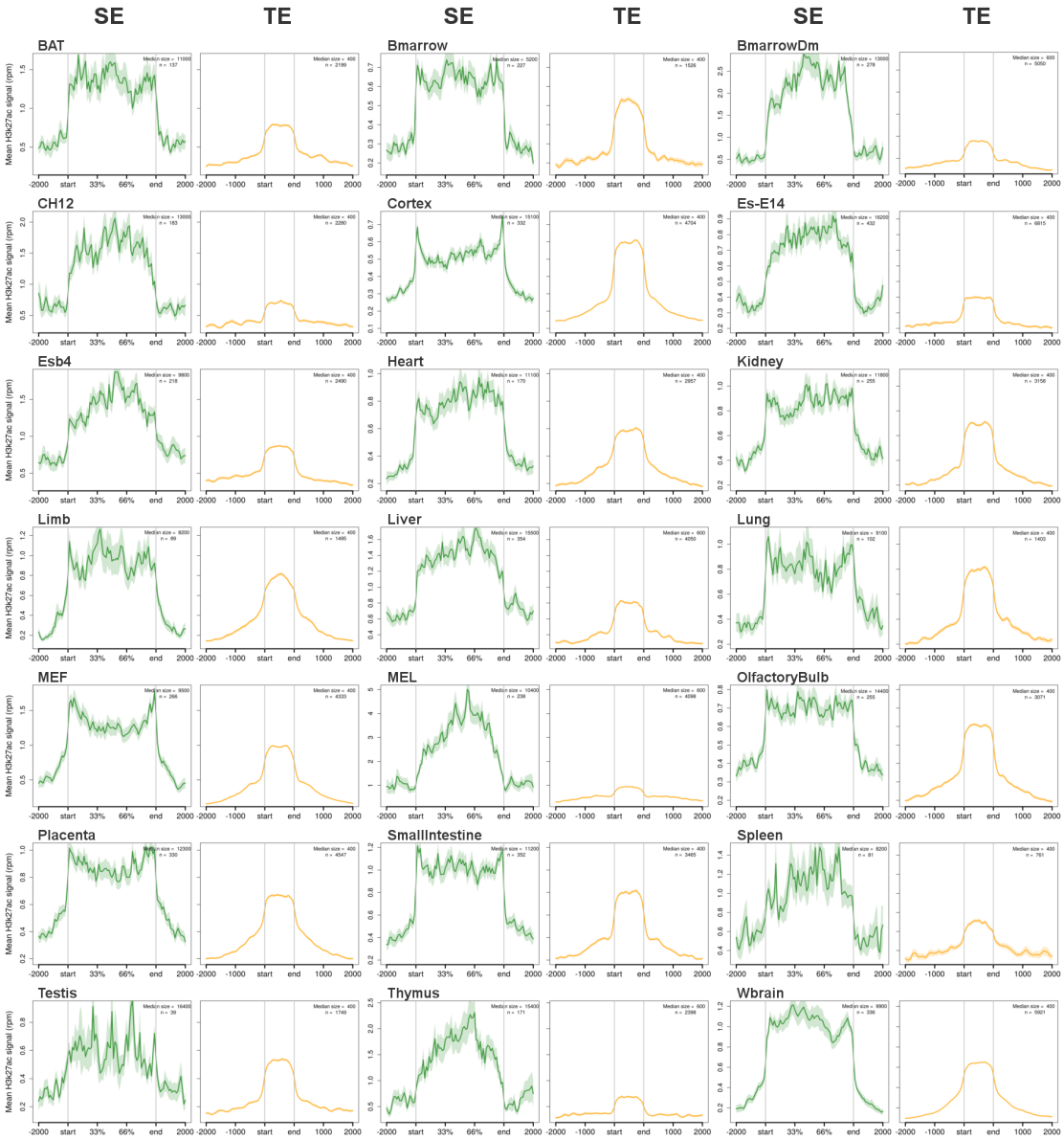


Fig. A.3 H3K27ac enrichment within SEs and TEs in 22 mouse tissues. Metagene profiles displaying mean H3K27ac ChIP-seq signal across all the SEs and TEs in each tissue. The profiles are centred on the enhancer region and the width represents the median length of the enhancers. An additional 2 kb region flanking each enhancer is also shown.

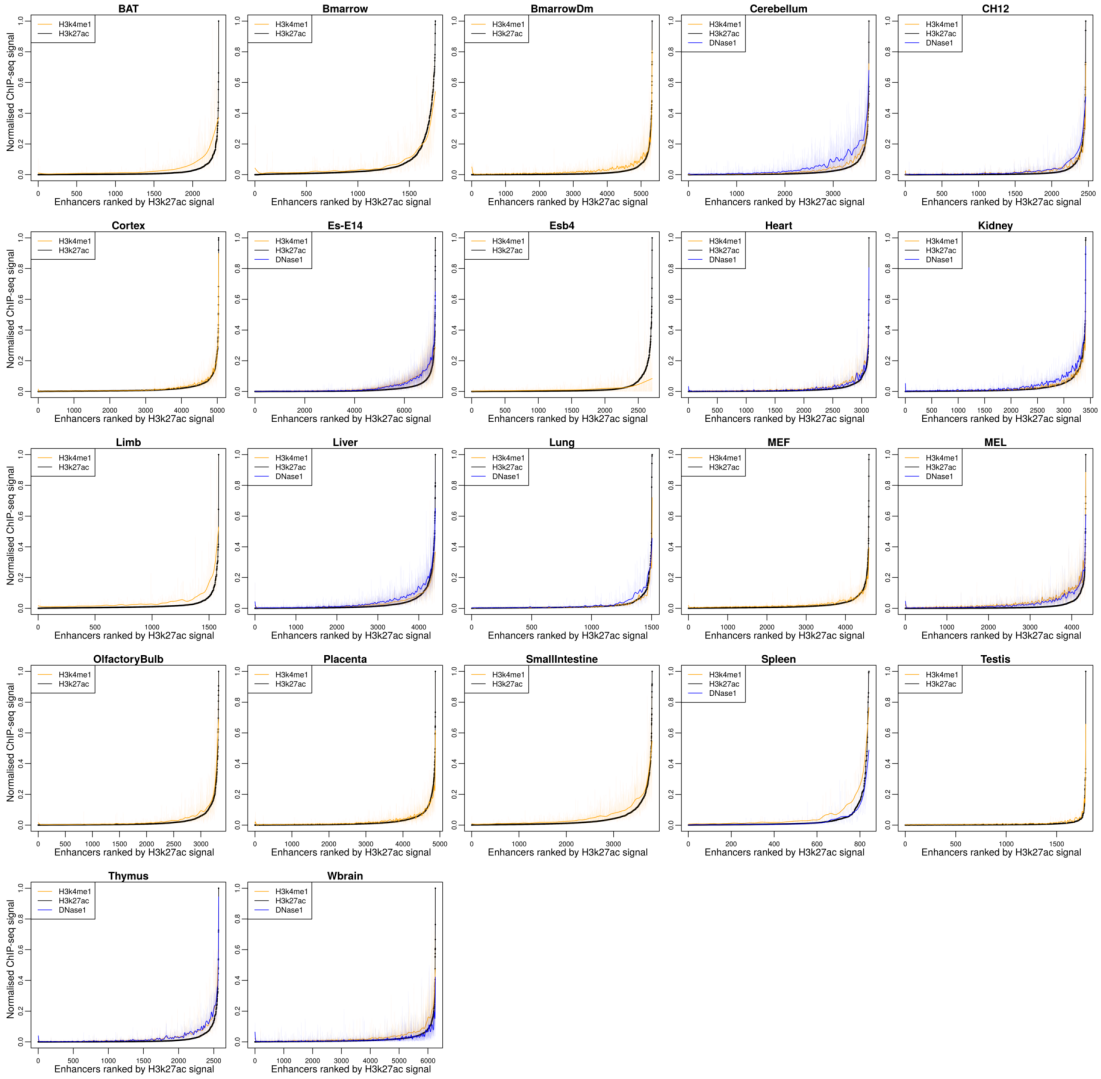


Fig. A.4 Comparison of H3K4me1, H3K27ac and DNaseI signal across stitched cohesive units. Distribution of H3K4me1, H3K27ac and DNaseI signal across stitched enhancers. The plot was normalised by dividing the input-subtracted ChIP-seq signal for each enhancer by the maximum ChIP-seq signal detected in each feature. The stitched enhancers for each feature on the x-axis are ranked according to the H3K27ac ChIP-seq signal. Please note that DNase-seq data was not available for all tissues.

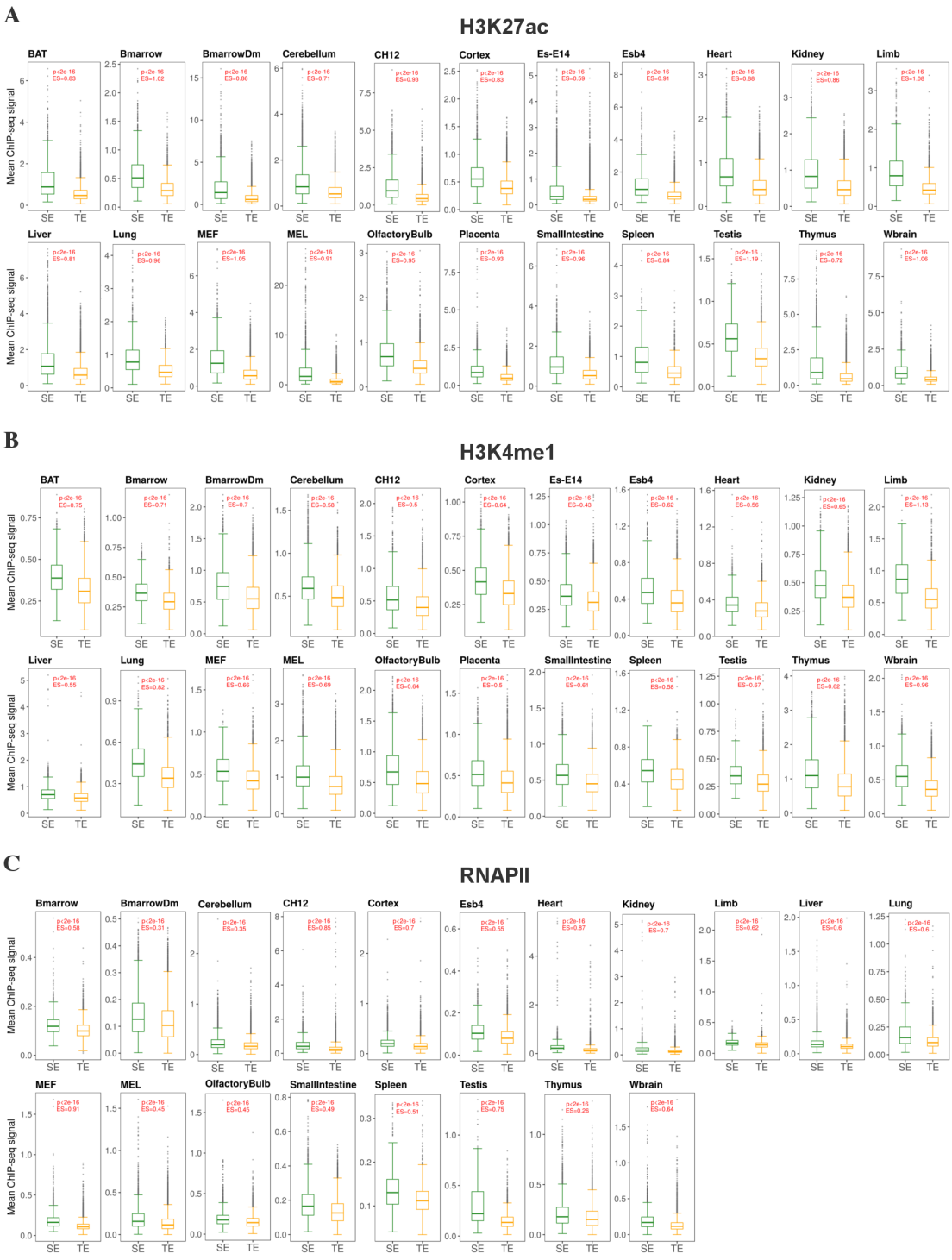


Fig. A.5 Chromatin activity within SE and TE constituent enhancers. Comparison of (A) H3K27ac, (B) H3K4me1 and (C) pol II ChIP-seq signal between SE and TE constituent enhancers in every tissue. Please note that pol II ChIP-seq data was not available for all tissues. p: p-values from Wilcoxon Rank Sum Test; ES: effect size.

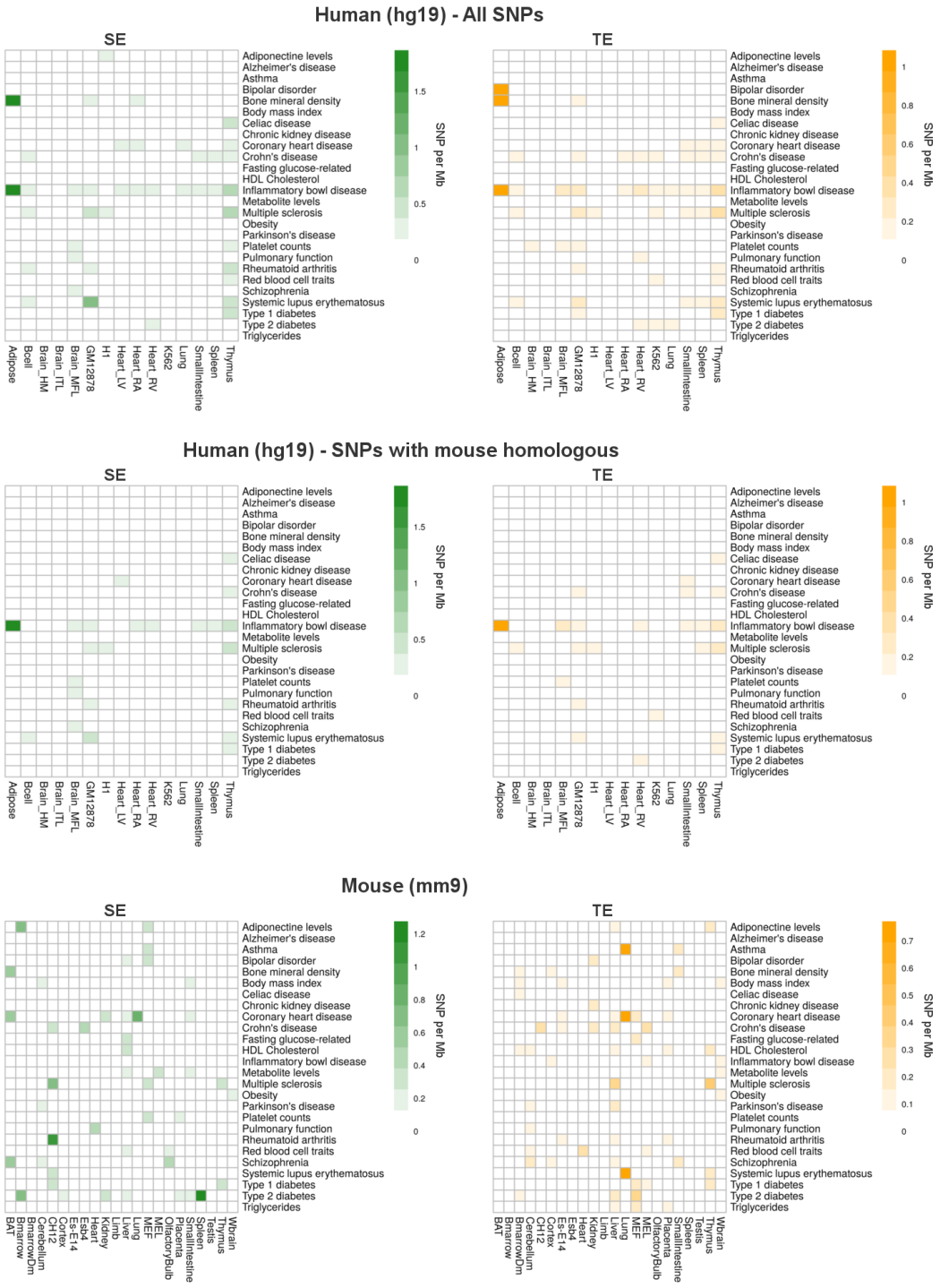


Fig. A.6 Enrichment of DA-SNPs from GWASs in SE and TE domains of human and mouse genomes. Heatmaps displaying the density of DA-SNPs occurring within SEs and TEs. The SNP density is represented as number of DA-SNPs occurring per Mb of enhancer.

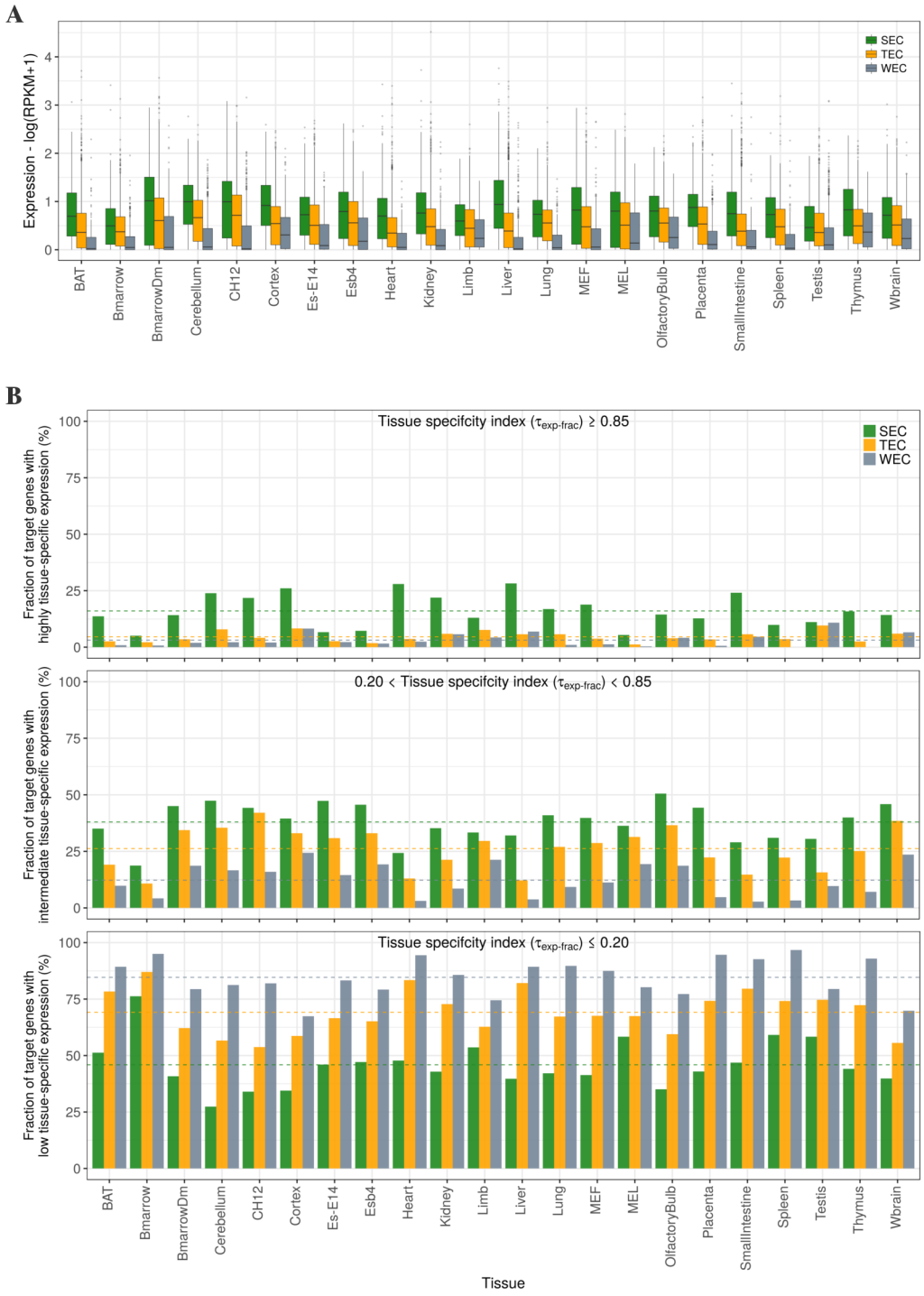


Fig. A.7 Effect of enhancer activity on target gene expression. (A) Box plot displaying the expression of genes in each enhancer class. (B) Bar plots showing the proportion of genes with high tissue-specific expression ($\tau_{exp-frac} \geq 0.85$), intermediate tissue-specific expression ($0.20 < \tau_{exp-frac} < 0.85$) and low tissue-specific expression ($\tau_{exp-frac} \leq 0.20$) in each enhancer class. The dotted lines show the mean across all the tissues.

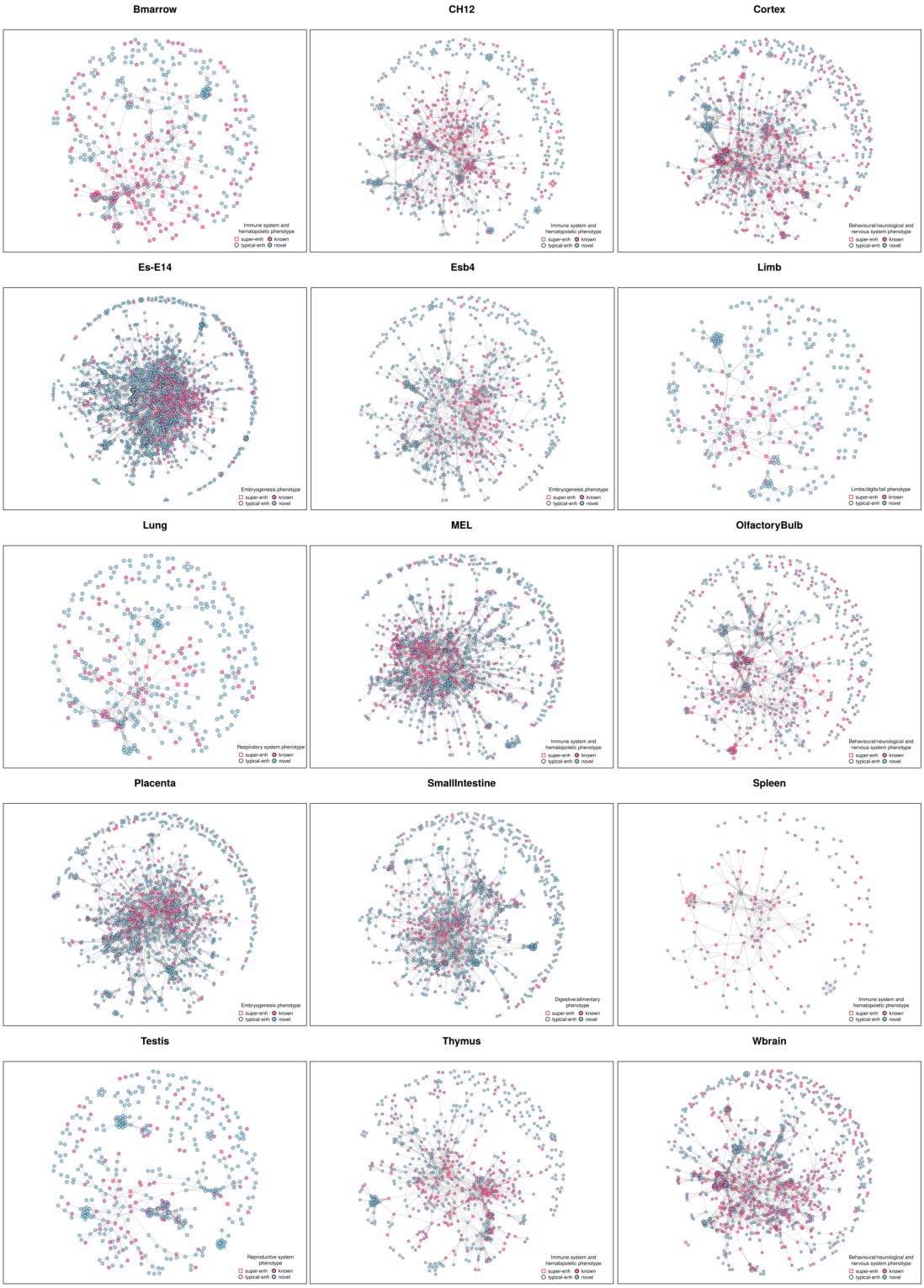


Fig. A.8 PPI maps of enhancer associated genes. Nodes in each network represents enhancer associated genes and edges represent a potential PPI between them. Genes associated with tissue-type relevant phenotypes are highlighted in pink and the shape of the node displays SE and TE associated genes (squares: SEC, circles: TEC).

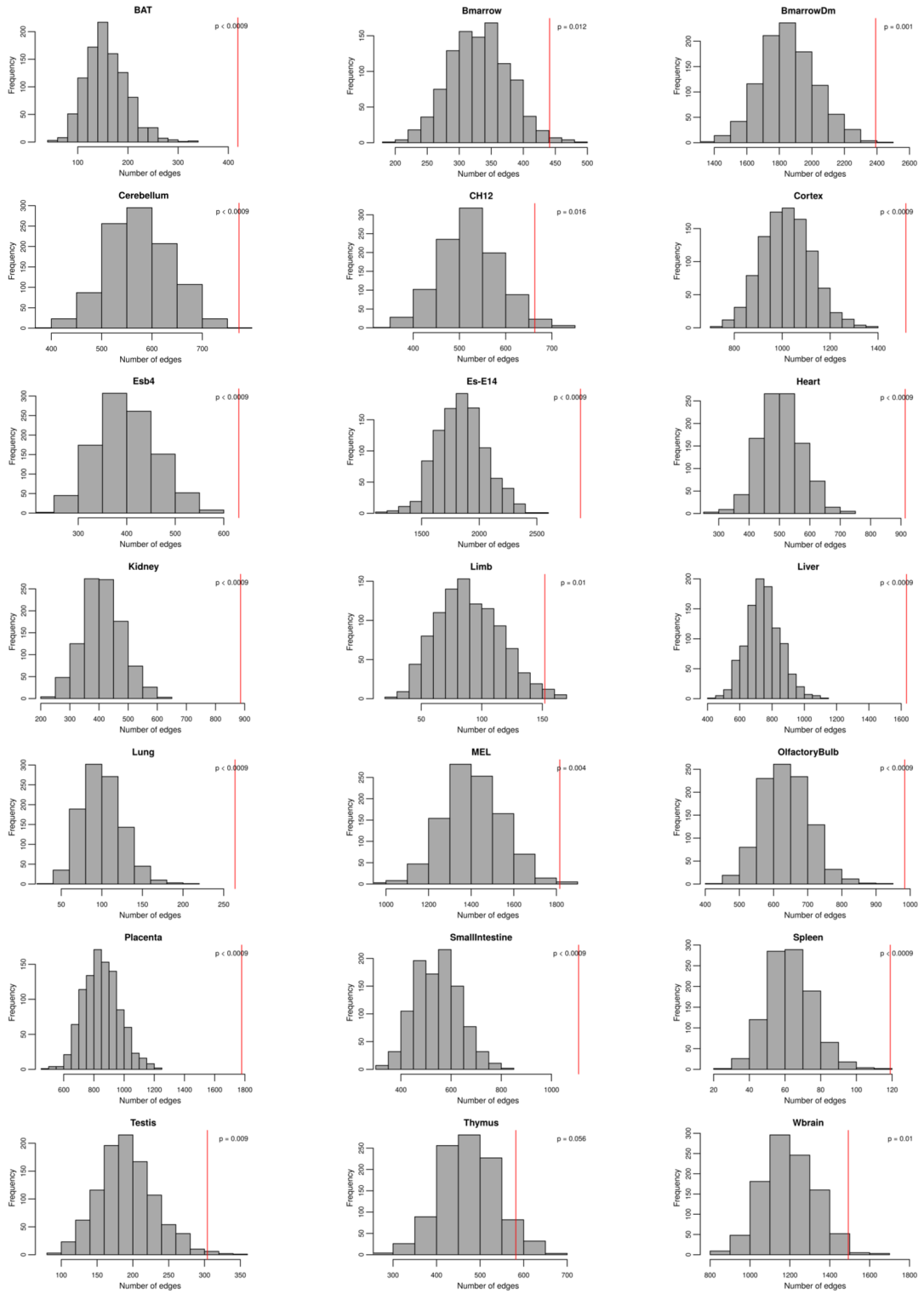


Fig. A.9 PPI network simulations. Histograms showing the PPI simulations performed for every tissue. The PPI networks were simulated 1,000 times by adding randomly selected protein-coding genes (equal to the number of phenotype associated genes in each tissue) to the network and counting their edges with the genes labelled as 'novel'. The red vertical line shows the observed number of edges between novel and phenotype associated genes and the grey distribution (obtained from permutation) shows the number of edges between novel and randomly added genes. The p-value shown was calculated by dividing the number of permuted values larger than the observed value by the total number of items in the permutation distribution.

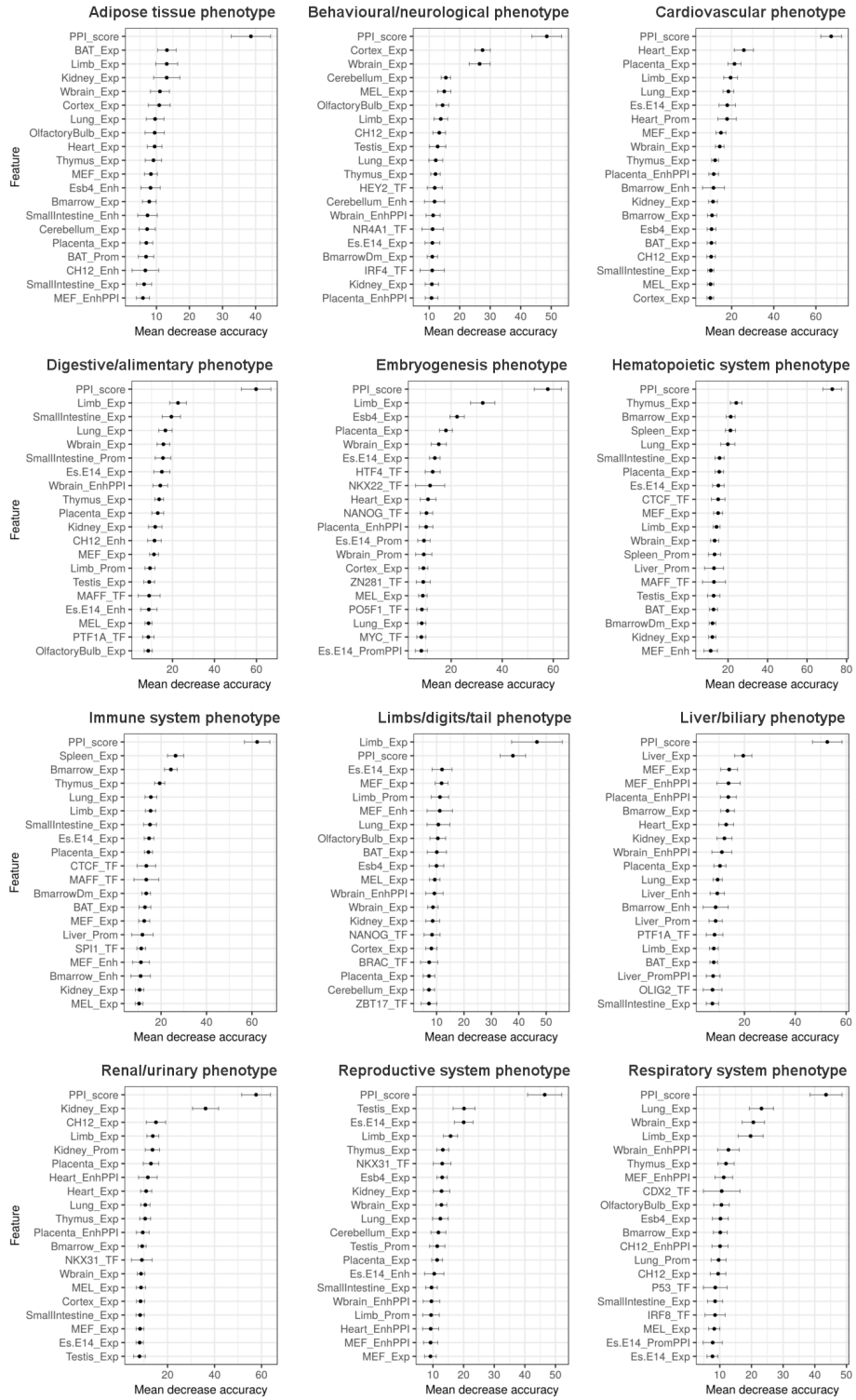


Fig. A.10 Feature importance of random forest classifiers used to predict gene-phenotype associations. Feature importance chart of the best performing model (Exp+PPI+TSRE+TSRE_PPI+TF) showing the top 20 predictor variables with the greatest contribution towards predicting gene-phenotype associations. Exp: expression; Enh: enhancer; Prom: promoter; TF: transcription factor.

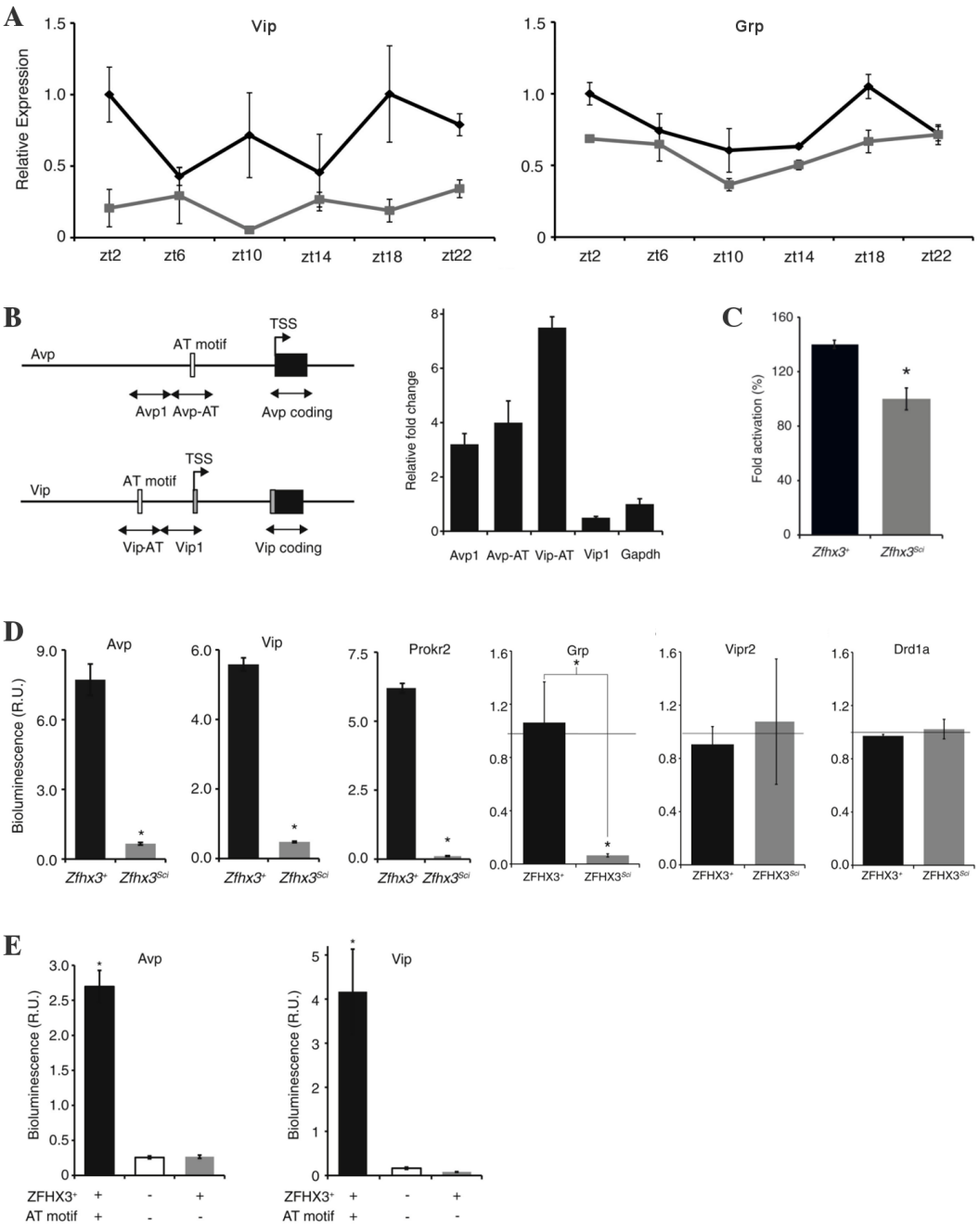


Fig. A.11 Experimental data related to the *Zfhx3^{Sci}* mutation. (A) Comparison of *Vip* and *Grp* mRNA expression in the SCN of *Zfhx3$^{Sci/+}$* (grey lines) and *Zfhx3$^{+/+}$* (black lines) across multiple time points (for both *Vip* and *Grp*: $n = 4$, $p < 0.05$, ANOVA). (B) Left panel shows a schematic representation of the regions used to design the primers spanning the AT motif and its adjacent regions. Right panel shows the results from quantitative ChIP in the SCN (time point: ZT3, $n = 3$). The bars display fold change with respect to the input (corresponding coding region of the gene). (C) Bar plot displaying the ability of *Zfhx3* to activate the AT motif in vitro ($n = 7$, $p < 0.05$, t-test). *Zfhx3^{Sci}*: *Zfhx3* with the *Sci* mutation ; *Zfhx3^{+}*: *Zfhx3* without the *Sci* mutation. (D) *In vitro* transcriptional activation of module 1 gene promoters by *Zfhx3* via the AT motif ('*' denotes $p < 0.05$, t-test). (E) Comparison of transcriptional activation by *Zfhx3* via the intact AT motif (+) and the disrupted AT motif (-). The three most conserved residues of the AT motif were mutated to disrupt the motif ($p < 0.05$, t-test). (Error bars show SEM)

Appendix B

Supplementary tables

Table B.1 GO enrichment analysis of SE associated genes.

Category	ID	Name	FDR B&H	# Genes
GO: MF	GO:0000981	RNA polymerase II transcription factor activity, sequence-specific DNA binding	6.64E-10	195
GO: MF	GO:0000982	transcription factor activity, RNA polymerase II proximal promoter sequence-specific DNA binding	4.01E-09	119
GO: MF	GO:0001228	transcriptional activator activity, RNA polymerase II transcription regulatory region sequence-specific DNA binding	3.35E-08	114
GO: MF	GO:0019904	protein domain specific binding	4.99E-08	198
GO: MF	GO:0022857	transmembrane transporter activity	2.10E-06	250
GO: MF	GO:0001227	transcriptional repressor activity	8.06E-06	63
GO: MF	GO:0098772	molecular function regulator	9.28E-06	330
GO: MF	GO:0044212	transcription regulatory region DNA binding	1.11E-05	213
GO: MF	GO:0003705	transcription factor activity, RNA polymerase II distal enhancer sequence-specific binding	1.87E-05	41
GO: MF	GO:0001012	RNA polymerase II regulatory region DNA binding	2.87E-05	163
GO: BP	GO:2000026	regulation of multicellular organismal development	9.95E-26	524
GO: BP	GO:0045595	regulation of cell differentiation	9.33E-23	470
GO: BP	GO:0022008	neurogenesis	8.17E-20	442
GO: BP	GO:0051094	positive regulation of developmental process	2.29E-17	366
GO: BP	GO:0060284	regulation of cell development	1.92E-16	287
GO: BP	GO:0051960	regulation of nervous system development	2.33E-15	259
GO: BP	GO:0048646	anatomical structure formation involved in morphogenesis	2.44E-15	352
GO: BP	GO:0006928	movement of cell or subcellular component	2.56E-15	476
GO: BP	GO:0009719	response to endogenous stimulus	2.67E-15	446
GO: BP	GO:0006811	ion transport	7.83E-15	420
GO: BP	GO:0072358	cardiovascular system development	1.11E-14	296
GO: BP	GO:0072359	circulatory system development	1.11E-14	296
GO: BP	GO:0009790	embryo development	1.75E-14	312
GO: BP	GO:0009891	positive regulation of biosynthetic process	1.97E-14	487
GO: BP	GO:0051049	regulation of transport	2.21E-14	490
GO: BP	GO:0010628	positive regulation of gene expression	2.56E-14	472
GO: BP	GO:0016477	cell migration	4.76E-14	346

Listed are the most enriched GO terms (MF: molecular function, BP: biological process) amongst SE associated genes. The GO enrichment analysis was performed using ToppGene. The associated FDR values were calculated using Benjamini-Hochberg method.

Table B.2 GO enrichment analysis of TE associated genes.

Category	ID	Name	FDR B&H	# Genes
GO: MF	GO:0017076	purine nucleotide binding	3.33E-06	1122
GO: MF	GO:0019899	enzyme binding	3.33E-06	1149
GO: MF	GO:0032549	ribonucleoside binding	3.33E-06	1093
GO: MF	GO:0001882	nucleoside binding	3.33E-06	1096
GO: MF	GO:0044877	protein-containing complex binding	3.33E-06	976
GO: MF	GO:0005524	ATP binding	6.63E-06	888
GO: MF	GO:0004672	protein kinase activity	2.01E-05	412
GO: MF	GO:0098772	molecular function regulator	8.07E-03	835
GO: MF	GO:0003713	transcription coactivator activity	3.67E-02	196
GO: BP	GO:0007010	cytoskeleton organization	9.32E-09	740
GO: BP	GO:0034613	cellular protein localization	9.45E-07	1013
GO: BP	GO:0010256	endomembrane system organization	1.78E-06	381
GO: BP	GO:0032989	cellular component morphogenesis	4.15E-06	873
GO: BP	GO:0015031	protein transport	1.03E-05	1162
GO: BP	GO:0051640	organelle localization	1.54E-05	335
GO: BP	GO:0044248	cellular catabolic process	1.55E-05	1092
GO: BP	GO:0000902	cell morphogenesis	3.36E-05	811
GO: BP	GO:1902531	regulation of intracellular signal transduction	3.36E-05	1062
GO: BP	GO:0044093	positive regulation of molecular function	7.02E-05	1122
GO: BP	GO:0022008	neurogenesis	1.29E-04	970
GO: BP	GO:0065003	protein-containing complex assembly	1.47E-04	1080
GO: BP	GO:0007049	cell cycle	2.74E-04	1044
GO: BP	GO:0070925	organelle assembly	3.30E-04	410
GO: BP	GO:0006629	lipid metabolic process	4.30E-04	823
GO: BP	GO:0048858	cell projection morphogenesis	4.65E-04	554
GO: BP	GO:0022402	cell cycle process	7.79E-04	825
GO: BP	GO:0016050	vesicle organization	7.95E-04	232
GO: BP	GO:0048699	generation of neurons	7.95E-04	905
GO: BP	GO:0006928	movement of cell or subcellular component	8.00E-04	1103
GO: BP	GO:0009894	regulation of catabolic process	8.04E-04	479

Listed are the most enriched GO terms (MF: molecular function, BP: biological process) amongst TE associated genes. The GO enrichment analysis was performed using ToppGene. The associated FDR values were calculated using Benjamini-Hochberg method.

Table B.3 Mammalian phenotype and human disease annotations enriched in enhancer classes.

	SEC			TEC		
Tissue	**N**	**Mouse Phenotypes**	**Disease**	**N**	**Mouse Phenotypes**	**Disease**
BAT	137	○ abnormal brown adipose tissue morphology, 10^{-3} ○ impaired adaptive thermogenesis, 10^{-3}	○ Fatty Liver, 10^{-3} ○ Strawberry nevus of skin, 10^{-3}	2199	○ abnormal cardiovascular system physiology, 10^{-9} ○ muscle phenotype, 10^{-7}	○ Diabetes Mellitus, Non-Insulin-Dependent, 10^{-8} ○ Obesity, 10^{-7}
BmarrowDm	278	○ abnormal adaptive immunity, 10^{-8} ○ abnormal immune cell physiology, 10^{-8}	○ Precursor Cell Lymphoblastic Leukemia Lymphoma, 10^{-3} ○ Autoimmune diseases, 10^{-3}	5050	○ abnormal professional antigen presenting cell morphology, 10^{-4} ○ abnormal macrophage morphology, 10^{-3}	○ Rheumatoid Arthritis, 10^{-7} ○ Diabetes Mellitus, Non-Insulin-Dependent, 10^{-7}
Bmarrow	227	○ abnormal neutrophil physiology, 10^{-3} ○ abnormal innate immunity, 10^{-3}	○ Neutropenia chronic, 10^{-2} ○ Sepsis, 10^{-2}	1526	○ abnormal blood cell physiology, 10^{-9} ○ abnormal immune cell physiology, 10^{-8}	○ Coronary heart disease, 10^{-6} ○ Pneumonia, 10^{-4}
Cerebellum	237	○ Impaired coordination, 10^{-8} ○ abnormal synaptic transmission, 10^{-7}	○ -	3504	○ abnormal synaptic transmission, 10^{-11} ○ abnormal nervous system physiology, 10^{-8}	◦ Bipolar Disorder, 10^{-7} ○ Unipolar Depression, 10^{-5}
CH12	183	○ abnormal B cell physiology, 10^{-12} ○ abnormal IgG level, 10^{-11}	○ Hodgkin Disease, 10^{-12} ○ Leukemogenesis, 10^{-12}	2251	◦ abnormal adaptive immunity, 10^{-9} ○ abnormal immune cell physiology, 10^{-9}	○ Leukemia, T-Cell, 10^{-5} ○ Adult T-Cell Lymphoma/Leukemia, 10^{-5}
Cortex	332	○ abnormal nervous system physiology, 10^{-13} ○ abnormal synaptic transmission, 10^{-13}	○ Schizophrenia, 10^{-9} ○ Unipolar Depression, 10^{-7}	4704	◦ abnormal synaptic transmission, 10^{-18} ○ abnormal nervous system physiology, 10^{-16}	○ Schizophrenia, 10^{-18} ○ Bipolar Disorder, 10^{-11}
Esb4	218	○ decreased primordial germ cell number, 10^{-2} ○ embryo phenotype, 10^{-2}	○ Seminoma, 10^{-2} ○ Mammary Neoplasms, 10^{-2}	2490	○ abnormal facial morphology, 10^{-4} ○ abnormal head morphology, 10^{-4}	○ Malignant neoplasm of ovary, 10^{-6} ○ Neuroblastoma, 10^{-5}
Es-E14	432	○ split vertebrae, 10^{-2}	○ -	6694	◦ abnormal nervous system development, 10^{-4} ○ abnormal embryo morphology, 10^{-3}	○ Autosomal recessive predisposition, 10^{-10} ○ Central neuroblastoma, 10^{-6}
Heart	170	○ abnormal muscle contractility, 10^{-16} ○ abnormal muscle fiber morphology, 10^{-16}	○ Left Ventricular Noncompaction, 10^{-14} ○ Cardiomyopathy, Dilated, 10^{-14}	2957	◦ muscle phenotype, 10^{-12} ○ abnormal cardiovascular system physiology, 10^{-9}	◦ Ventricular arrhythmia, 10^{-6} ○ Adenoma, 10^{-6}
Kidney	225	○ abnormal kidney morphology, 10^{-3} ○ abnormal renal/urinary system morphology, 10^{-2}	◦ Renal cyst, 10^{-4} ○ Polycystic Kidney Diseases, 10^{-3}	3158	○ renal/urinary system phenotype, 10^{-7} ○ abnormal urine homeostasis, 10^{-5}	◦ Malignant tumor of colon, 10^{-6} ○ Mammary Neoplasms, 10^{-5}
Limb	89	○ abnormal neurocranium morphology, 10^{-4} ○ abnormal basicranium morphology, 10^{-4}	○ Mixed Salivary Gland Tumor, 10^{-2} ○ Limited elbow extension, 10^{-2}	1495	○ abnormal craniofacial morphology, 10^{-15} ○ abnormal limb morphology, 10^{-14}	◦ Congenital Abnormality, 10^{-7} ○ Ventricular Septal Defects, 10^{-7}
Liver	354	○ liver/biliary system phenotype, 10^{-10} ○ abnormal lipid homeostasis, 10^{-10}	○ Non-alcoholic Fatty Liver Disease, 10^{-11} ○ Steatohepatitis, 10^{-11}	4050	◦ abnormal hepatobiliary system physiology, 10^{-9} ○ abnormal liver physiology, 10^{-8}	◦ Liver neoplasms, 10^{-12} ○ Fatty Liver, 10^{-12}
Lung	102	○ abnormal pulmonary alveolus morphology, 10^{-2} ○ abnormal pulmonary acinus morphology, 10^{-2}	○ Synovial Cyst, 10^{-2} ○ Myxoid cyst, 10^{-2}	1403	○ abnormal cardiovascular development, 10^{-6} ○ abnormal vascular development, 10^{-6}	◦ Congenital Abnormality, 10^{-5} ○ Hypertensive disease, 10^{-4}
MEF	266	○ abnormal vascular development, 10^{-5} ○ abnormal tracheal cartilage morphology, 10^{-5}	○ Epithelial ovarian cancer, 10^{-7} ○ Liver Cirrhosis, Experimental, 10^{-7}	4333	○ craniofacial phenotype, 10^{-14} ○ perinatal lethality, 10^{-10}	◦ Endometriosis, 10^{-11} ◦ Mammary Neoplasms, 10^{-11}
MEL	238	○ -	○ Ecchymosis, 10^{-4} ○ Increased tendency to bruise, 10^{-4}	4098	○ -	○ Schizophrenia, 10^{-5} ○ Epilepsy, 10^{-4}
OlfactoryBulb	255	○ abnormal CNS synaptic transmission, 10^{-4} ○ reduced long term potentiation, 10^{-4}	○ Schizophrenia, 10^{-9} ○ Abnormal behavior, 10^{-4}	3071	○ abnormal synaptic transmission, 10^{-13} ○ abnormal nervous system physiology, 10^{-12}	○ Bipolar Disorder, 10^{-7} ○ Schizophrenia, 10^{-6}
Placenta	330	◦ embryonic lethality during organogenesis, incomplete penetrance, 10^{-5} ○ embryonic lethality, 10^{-4}	○ Malignant neoplasm of pancreas, 10^{-3} ○ Pancreatic carcinoma, 10^{-2}	4547	○ lethality throughout fetal growth and development, 10^{-4} ○ embryo phenotype, 10^{-4}	○ Malignant neoplasm of pancreas, 10^{-7} ◦ Pancreatic carcinoma, 10^{-6}
SmallIntestine	352	○ abnormal digestive system physiology, 10^{-4} ○ abnormal large intestine morphology, 10^{-4}	◦ Malignant tumor of colon, 10^{-7} ○ Colorectal cancer metastatic, 10^{-6}	3465	○ abnormal lipid homeostasis, 10^{-3} ○ abnormal lipid level, 10^{-2}	◦ Liver neoplasms, 10^{-8} ○ Malignant tumor of colon, 10^{-6}
Spleen	81	○ abnormal IgM level, 10^{-8} ○ abnormal definitive hematopoiesis, 10^{-8}	○ Chronic Lymphocytic Leukemia, 10^{-6} ○ Precursor Cell Lymphoblastic Leukemia Lymphoma, 10^{-5}	761	○ abnormal leukocyte physiology, 10^{-15} ◦ abnormal adaptive immunity, 10^{-15}	◦ Autoimmune Diseases, 10^{-8} ○ Diffuse Large B-Cell Lymphoma, 10^{-8}
Testis	39	○ abnormal Leydig cell morphology, 10^{-2}	○ Male infertility, 10^{-2} ○ Testicular anomalies with or without congenital heart disease, 10^{-2}	1749	○ -	○ -
Thymus	171	○ abnormal positive T cell selection, 10^{-10} ◦ decreased T cell number, 10^{-10}	○ Lymphoma, 10^{-3} ○ Lupus Erythematosus, Systemic, 10^{-3}	2398	○ abnormal lymphocyte cell number, 10^{-9} ○ abnormal T cell number, 10^{-9}	◦ B-Cell Lymphomas, 10^{-6} ○ Leukemogenesis, 10^{-5}
Wbrain	336	○ abnormal amacrine cell number, 10^{-4} ○ abnormal axon fasciculation, 10^{-4}	○ Seizures, Focal, 10^{-2} ○ Substance-Related Disorders, 10^{-2}	5921	◦ abnormal neuron morphology, 10^{-19} ○ abnormal somatic nervous system development, 10^{-16}	◦ Schizophrenia, 10^{-19} ○ Bipolar Disorder, 10^{-12}

Listed are the most enriched mammalian phenotypes and human disease annotation terms in each tissue. The associated FDR values are reported next to the enriched terms and was calculated using Benjamini-Hochberg method. N displays the number of enhancer associated genes in each group.

Table B.4 Mammalian phenotype and human disease annotations enriched in the WEC.

		WEC	
Tissue	**N**	**Mouse Phenotypes**	**Disease**
BAT	214	○ -	○ Otofaciocervical Syndrome, 10^{-3}
BmarrowDm	939	○ -	○ Autism Spectrum Disorders, 10^{-3} ○ Mental Retardation, X-Linked, 10^{-3}
Bmarrow	368	○ -	○ -
Cerebellum	1202	○ abnormal innervation, 10^{-5} ○ abnormal synaptic transmission, 10^{-4}	○ Autistic Disorder, 10^{-10} ○ Schizophrenia,10^{-6}
CH12	1117	○ abnormal cartilage morphology, 10^{-2} ○ perinatal lethality, 10^{-2}	○ Mental Depression, 10^{-4} ○ Alcoholic Intoxication, Chronic, 10^{-4}
Cortex	1006	○ abnormal synaptic transmission, 10^{-14} ○ abnormal nervous system physiology, 10^{-13}	○ Bipolar Depression, 10^{-12} ○ Schizophrenia, 10^{-11}
Esb4	391	○ increased neurotransmitter release, 10^{-3}	○ Autistic Disorder, 10^{-2}
Es-E14	687	○ perinatal lethality, 10^{-6} ○ neonatal lethality, 10^{-5}	○ Autism Spectrum Disorders, 10^{-3} ○ Craniofacial Abnormalities, 10^{-2}
Heart	335	○ abnormal muscle contractility, 10^{-4} ○ increased heart ventricle size, 10^{-3}	○ -
Kidney	449	○ -	◦ Alcoholic Intoxication, Chronic, 10^{-2} ○ Autism Spectrum Disorders, 10^{-2}
Limb	87	○ abnormal palatal shelf fusion at midline, 10^{-3} ○ cleft hard palate, 10^{-2}	○ -
Liver	922	○ -	○ Attention deficit hyperactivity, 10^{-5} ○ Autistic Disorder, 10^{-5}
Lung	286	○ abnormal olfactory lobe morphology, 10^{-3} ○ abnormal nervous system physiology, 10^{-2}	○ -
MEF	601	○ muscle phenotype, 10^{-3} ○ abnormal limb bone morphology, 10^{-3}	○ Ventricular Septal Defects, 10^{-2} ○ Alcoholic Intoxication, Chronic, 10^{-2}
MEL	621	○ -	○ Autistic Disorder, 10^{-4} ○ Mental Retardation, X-Linked, 10^{-4}
OlfactoryBulb	360	○ abnormal motor coordination/balance, 10^{-4} ○ abnormal nervous system physiology, 10^{-3}	○ Autistic Disorder, 10^{-4} ○ Fragile X Syndrome, 10^{-3}
Placenta	975	◦ abnormal neuron differentiation, 10^{-3} ○ abnormal respiratory system physiology, 10^{-3}	○ Central neuroblastoma, 10^{-7} ○ Alzheimer's Disease, 10^{-7}
SmallIntestine	310	○ abnormal gallbladder physiology, 10^{-3} ○ postnatal lethality, 10^{-2}	◦ Obesity, 10^{-2}
Spleen	170	○ abnormal semicircular canal morphology, 10^{-3} ○ abnormal otolith organ morphology, 10^{-2}	○ Hypoplastic cochlea, 10^{-2} ○ Congenital Abnormality, 10^{-2}
Testis	864	○ abnormal neuron morphology, 10^{-3} ○ abnormal synaptic transmission, 10^{-3}	○ Autistic Disorder, 10^{-4} ○ Craniofacial Abnormalities, 10^{-3}
Thymus	412	○ -	○ -
Wbrain	776	○ abnormal neuron differentiation, 10^{-10} ○ abnormal nervous system development, 10^{-8}	○ Schizophrenia, 10^{-6} ○ Autism Spectrum Disorders, 10^{-6}

Listed are the most enriched mammalian phenotypes and human disease annotation terms in the WEC. The associated FDR values are reported next to the enriched terms and was calculated using Benjamini-Hochberg method. N displays the number of weak-enhancer associated genes in each group.

Table B.5 Disease annotation terms enriched amongst genes associated with nervous system phenotype in the mouse.

Name	Source	P-value	FDR B&H	Genes from Input	Genes in Annotation
Schizophrenia	DisGeNET Curated	1.06E-154	2.05E-150	750	1561
Seizures	DisGeNET Curated	1.22E-148	1.17E-144	551	982
Alzheimer's Disease	DisGeNET Curated	2.61E-138	1.67E-134	803	1825
Epilepsy	DisGeNET Curated	6.45E-135	3.10E-131	526	962
Neuroblastoma	DisGeNET Curated	1.04E-127	4.00E-124	745	1689
Central neuroblastoma	DisGeNET BeFree	5.04E-126	1.62E-122	725	1631
Intellectual Disability	DisGeNET Curated	1.42E-118	3.91E-115	559	1131
Depressive disorder	DisGeNET Curated	1.02E-112	2.45E-109	403	697
Mental Retardation	DisGeNET Curated	4.82E-107	1.03E-103	461	885
Bipolar Disorder	DisGeNET Curated	1.99E-104	3.83E-101	404	730
Impaired cognition	DisGeNET Curated	5.68E-95	9.93E-92	354	625
Mental Depression	DisGeNET Curated	3.48E-94	5.58E-91	327	554
Low intelligence	DisGeNET Curated	2.26E-89	2.90E-86	355	649
Dull intelligence	DisGeNET Curated	2.26E-89	2.90E-86	355	649
Poor school performance	DisGeNET Curated	2.26E-89	2.90E-86	355	649
Mental deficiency	DisGeNET Curated	7.28E-89	8.75E-86	355	651
Autistic Disorder	DisGeNET Curated	3.58E-84	4.05E-81	346	644
Autosomal recessive predisposition	DisGeNET Curated	5.64E-83	6.03E-80	592	1445
Congenital Abnormality	DisGeNET Curated	1.82E-82	1.84E-79	330	605
Hyperactive behaviour	DisGeNET Curated	1.03E-81	9.85E-79	416	866
Neurodegenerative Disorders	DisGeNET Curated	4.39E-81	4.02E-78	364	710
Abnormal behaviour	DisGeNET Curated	8.74E-81	7.64E-78	238	365
Glioblastoma	DisGeNET Curated	4.07E-79	3.40E-76	696	1851
Malignant neoplasm of ovary	DisGeNET Curated	5.12E-76	4.11E-73	705	1911
nervous system disorder	DisGeNET Curated	7.85E-72	6.04E-69	250	423
Mammary Neoplasms	DisGeNET Curated	7.20E-71	5.33E-68	704	1953
Amyloidosis	DisGeNET Curated	3.28E-70	2.33E-67	412	916
Global developmental delay	DisGeNET Curated	1.90E-69	1.30E-66	361	757
Obesity	DisGeNET Curated	2.03E-69	1.34E-66	661	1803
Major Depressive Disorder	DisGeNET Curated	5.63E-69	3.61E-66	276	505
Malignant neoplasm of pancreas	DisGeNET Curated	9.37E-69	5.81E-66	640	1730
Parkinson Disease	DisGeNET Curated	3.46E-68	2.08E-65	412	928
Hypertensive disease	DisGeNET Curated	8.35E-68	4.86E-65	466	1112
Pancreatic carcinoma	DisGeNET Curated	1.35E-67	7.66E-65	649	1774
Autism Spectrum Disorders	DisGeNET Curated	1.40E-66	7.72E-64	248	436
Unipolar Depression	DisGeNET Curated	2.87E-66	1.54E-63	241	418
Malignant tumour of colon	DisGeNET Curated	1.10E-64	5.69E-62	679	1915
Adenoma	DisGeNET Curated	1.78E-63	9.02E-61	414	964
Anxiety	DisGeNET Curated	3.42E-63	1.69E-60	242	432
Squamous cell carcinoma	DisGeNET Curated	3.77E-61	1.81E-58	602	1655
Astrocytoma	DisGeNET Curated	7.52E-61	3.53E-58	335	721
Brain Neoplasms	DisGeNET Curated	1.25E-60	5.73E-58	302	619
Muscle hypotonia	DisGeNET Curated	2.77E-60	1.24E-57	287	575
leukaemia	DisGeNET Curated	3.01E-60	1.32E-57	652	1854
Anxiety Disorders	DisGeNET Curated	1.35E-59	5.78E-57	219	382
Medulloblastoma	DisGeNET Curated	8.74E-59	3.65E-56	279	558
Psychotic Disorders	DisGeNET Curated	8.78E-58	3.59E-55	190	311
Huntington Disease	DisGeNET Curated	1.69E-57	6.75E-55	292	604
Amyotrophic Lateral Sclerosis	DisGeNET Curated	1.83E-57	7.18E-55	300	629
Mental and motor retardation	DisGeNET Curated	2.33E-57	8.78E-55	295	614

Table B.6 GO enrichment analysis of genes in module 4 of the *Zfhx3*$^{Sci/+}$ network.

Category	ID	Name	Corrected p-value	Genes
GO:BP	GO:0034368	protein-lipid complex remodelling	8.45E-05	APOA2, PLA2G7, APOA1
GO:BP	GO:0034369	plasma lipoprotein particle remodelling	8.45E-05	APOA2, PLA2G7, APOA1
GO:BP	GO:0002740	negative regulation of cytokine secretion involved in immune response	2.33E-03	APOA2, APOA1
GO:BP	GO:0002682	regulation of immune system process	4.78E-03	APOA2, PLA2G7, APOA1, RORA, LRRC17, H2-D1
GO:BP	GO:0042632	cholesterol homeostasis	5.79E-03	APOA2, APOA1, RORA
GO:BP	GO:0030300	regulation of intestinal cholesterol absorption	6.52E-03	APOA2, APOA1
GO:BP	GO:0018206	peptidyl-methionine modification	8.38E-03	APOA2, APOA1
GO:BP	GO:0010873	positive regulation of cholesterol esterification	8.38E-03	APOA2, APOA1
GO:BP	GO:0070508	cholesterol import	1.28E-02	APOA2, APOA1
GO:BP	GO:0018158	protein oxidation	1.28E-02	APOA2, APOA1
GO:BP	GO:0060192	negative regulation of lipase activity	1.81E-02	APOA2, APOA1
GO:BP	GO:0033700	phospholipid efflux	1.81E-02	APOA2, APOA1
GO:BP	GO:0043691	reverse cholesterol transport	2.44E-02	APOA2, APOA1
GO:BP	GO:0006656	phosphatidylcholine biosynthetic process	3.56E-02	APOA2, APOA1
GO:BP	GO:0097164	ammonium ion metabolic process	4.56E-02	APOA2, PLA2G7, APOA1
GO:MF	GO:0070653	high-density lipoprotein particle receptor binding	2.33E-04	APOA2, APOA1
GO:MF	GO:0032934	sterol binding	7.86E-04	APOA2, APOA1, RORA
GO:MF	GO:0060228	phosphatidylcholine-sterol O-acyltransferase activator activity	2.33E-03	APOA2, APOA1
GO:MF	GO:0034190	apolipoprotein receptor binding	4.89E-03	APOA2, APOA1
GO:MF	GO:0008035	high-density lipoprotein particle binding	8.38E-03	APOA2, APOA1
GO:MF	GO:0055102	lipase inhibitor activity	1.54E-02	APOA2, APOA1
GO:MF	GO:0017127	cholesterol transporter activity	3.56E-02	APOA2, APOA1

Listed are the most enriched GO terms (BP: biological process, MF: molecular function) amongst differentially expressed genes in module 4 of the *Zfhx3*$^{Sci/+}$ network. The GO enrichment analysis was performed using g:Profiler.